

Enzymatic Promiscuity

A Thesis
Presented to
The Division of Mathematics and Natural Sciences
Reed College

In Partial Fulfillment
of the Requirements for the Degree
Bachelor of Arts

Nelly Selem

May 2016

Approved for the Division
(Mathematics)

Francisco Barona Gomez

Acknowledgements

I want to thank a few people.

Preface

This is an example of a thesis setup to use the reed thesis document class.

Table of Contents

Introduction	1
Chapter 1: EvoMining	3
1.1 Genome DB	3
1.2 Natural Products DB	4
1.3 Reproducibility on EvoMining code	4
1.4 Otras estrategias para los clusters Argon context Idea	4
1.5 Inline code	5
1.6 CORASoN	5
1.7 Loading and exploring data	6
1.8 Additional resources	9
Chapter 2: PriA Family	11
2.1 Math	11
2.2 Chemistry 101: Symbols	11
2.2.1 Typesetting reactions	12
2.2.2 Other examples of reactions	12
2.3 Physics	12
2.4 Biology	12
Chapter 3: Archaea EvoMining Results	13
3.1 Tables	13
3.1.1 Expansions BoxPlot by metabolic family	14
3.2 Central pathway expansions	16
3.3 Genome Size correlations	17
3.3.1 Correlation between genome size and AntiSMASH products .	17
3.3.2 Correlation between genome size and Central pathway expansions	19
3.4 Natural products	23
3.4.1 Natural products recruitments from EvoMining heatplot . .	23
3.5 Archaeas AntiSMASH	25
3.5.1 AntisSMASH vs Central Expansions	27
3.6 Selected trees from EvoMining	30
3.7 Footnotes and Endnotes	31
3.8 Bibliographies	31
3.9 Anything else?	32

Chapter 4: Actinobacteria EvoMining Results	33
4.1 Tables	33
4.1.1 Expansions BoxPlot by metabolic family	34
4.2 Central pathway expansions	36
4.3 Genome Size correlations	40
4.3.1 Correlation between genome size and AntiSMASH products . .	40
4.3.2 Correlation between genome size and Central pathway expansions	42
4.4 Natural products	46
4.4.1 Natural products recruitments from EvoMining heatplot . . .	46
4.5 Actinos AntiSMASH	48
4.5.1 AntisSMASH vs Central Expansions	50
4.6 Selected trees from EvoMining	53
Chapter 5: Cyanobacteria EvoMining Results	55
5.1 Tables	55
5.1.1 Expansions BoxPlot by metabolic family	56
5.2 Central pathway expansions	58
5.3 Genome Size correlations	59
5.3.1 Correlation between genome size and AntiSMASH products .	59
5.3.2 Correlation between genome size and Central pathway expansions	61
5.4 Natural products	65
5.4.1 Natural products recruitments from EvoMining heatplot . . .	65
5.5 Cyanobacterias AntiSMASH	67
5.5.1 AntisSMASH vs Central Expansions	69
5.6 Selected trees from EvoMining	72
Conclusion	77
Appendix A: The First Appendix	79
Appendix B: The Second Appendix, for Fun	81
References	83

List of Tables

1.1	Maximum Delays by Airline	8
3.1	Families on Archaeabacteria	13
4.1	Correlation of Inheritance Factors for Parents and Child	33
5.1	Families on Cyanobacteria	55

List of Figures

2.1	Combustion of glucose	12
3.1	Expansions Boxplot	15
3.2	Archaeas Heatplot	16
3.3	Correlation between Archaeas genome size and antismash Natural products detection colored by Order	17
3.4	Correlation between Archaeas genome size and antismash Natural products detection grided by Order	18
3.5	Correlation between Archaeas genome size and central pathway expansions	19
3.6	Correlation between Archaeas genome size and central pathway expansions grided by order	20
3.7	Correlation between Archaeas Genome size vs Total central pathway expansion coloured by metabolic Family	21
3.8	Archaeas Recruitmens on central families coloured by kingdom	23
3.9	Archaeas Recruitmens on central families coloured by taxonomy	24
3.10	Archaeas Diversity	25
3.11	Archaeas Smash Taxonomical Diversity	26
3.12	Correlation between Archaeas central pathway expansions and anti-smash Natural products detection	27
3.13	Correlation between Archaeas central pathway expansions and anti-smash Natural products detection	28
3.14	Archaeas Natural products by family	29
3.15	Phosphoribosyl isomerase A EvoMiningtree	30
3.16	Phosphoribosyl isomerase other EvoMiningtree	30
3.17	Phosphoribosyl anthranilate isomerase EvoMiningtree	30
4.1	Expansions Boxplot	35
4.2	Actinobacterial Heatplot	37
4.3	Streptomyces Genomes expansions on PGA Aminoacids HeatPlot	39
4.4	Correlation between Actinos genome size and antismash Natural products detection colored by Order	40
4.5	Correlation between Actinos genome size and antismash Natural products detection grided by Order	41
4.6	Correlation between Actinos genome size and central pathway expansions	42

4.7	Correlation between Actinos genome size and central pathway expansions grided by order	43
4.8	Correlation between Actinos Genome size vs Total central pathway expansion coloured by metabolic Family	44
4.9	Actinos Recruitmens on central families coloured by kingdom	46
4.10	Actinos Recruitmens on central families coloured by taxonomy	47
4.11	Actinos Diversity	48
4.12	Actinos Smash Taxonomical Diversity	49
4.13	Correlation between Actinos central pathway expansions and antismash Natural products detection	50
4.14	Correlation between Actinos central pathway expnasions and antismash Natural products detection	51
4.15	Actinos Natural products by family	52
4.16	Enolase EvoMiningtree	53
4.17	Phosphoribosyl isomerase EvoMiningtree	53
4.18	Phosphoribosyl isomerase A EvoMiningtree	53
4.19	phosphoshikimate carboxyvinyltransferase EvoMiningtree	54
5.1	Expansions Boxplot	57
5.2	Cyanobacterial Heatplot	58
5.3	Correlation between genome size and antismash Natural products detection colored by Order	59
5.4	Correlation between genome size and antismash Natural products detection grided by Order	60
5.5	Correlation between genome size and central pathway expansions	61
5.6	Correlation between genome size and central pathway expansions grided by order	62
5.7	Correlation between Genome size vs Total central pathway expansion coloured by metabolic Family	63
5.8	Recruitmens on central families coloured by kingdom	65
5.9	Recruitmens on central families coloured by taxonomy	66
5.10	Diversity	67
5.11	Smash	68
5.12	Correlation between central pathway axpnasions and antismash Natural products detection	69
5.13	Correlation between central pathway axpnasions and antismash Natural products detection	70
5.14	Natural products by family	71
5.15	Phosphoribosyl isomerase EvoMiningtree	72
5.16	Phosphoglycerate dehydrogenase EvoMiningtree	72
5.17	Phosphoserine aminotransferase EvoMiningtree	72
5.18	Triosephosphate isomerase EvoMiningtree	73
5.19	glyceraldehyde3phosphate dehydrogenase EvoMiningtree	73
5.20	phosphoglycerate kinase EvoMiningtree	73
5.21	phosphoglycerate mutaseEvoMiningtree	74

5.22 enolase EvoMiningtree	74
5.23 Pyruvate kinase EvoMiningtree	74
5.24 Aspartate transaminase EvoMiningtree	75
5.25 Asparagine synthase EvoMiningtree	75
5.26 Aspartate kinase EvoMiningtree	75
5.27 Aspartate semialdehyde dehydrogenase EvoMiningtree	76
5.28 Homoserine dehydrogenase EvoMiningtree	76

Abstract

The preface pretty much says it all.
Second paragraph of abstract starts here.

Dedication

You can have a dedication here if you wish.

Introduction

Welcome to the *R Markdown* thesis template. This template is based on (and in many places copied directly from) the *L^AT_EX* template, but hopefully it will provide a nicer interface for those that have never used *T_EX* or *L^AT_EX* before. Using *R Markdown* will also allow you to easily keep track of your analyses in **R** chunks of code, with the resulting plots and output included as well. The hope is this *R Markdown* template gets you in the habit of doing reproducible research, which benefits you long-term as a researcher, but also will greatly help anyone that is trying to reproduce or build onto your results down the road.

Hopefully, you won't have much of a learning period to go through and you will reap the benefits of a nicely formatted thesis. The use of *L^AT_EX* in combination with *Markdown* is more consistent than the output of a word processor, much less prone to corruption or crashing, and the resulting file is smaller than a Word file. While you may have never had problems using Word in the past, your thesis is likely going to be about twice as large and complex as anything you've written before, taxing Word's capabilities. After working with *Markdown* and **R** together for a few weeks, we are confident this will be your reporting style of choice going forward.

Why use it?

R Markdown creates a simple and straightforward way to interface with the beauty of *L^AT_EX*. Packages have been written in **R** to work directly with *L^AT_EX* to produce nicely formatting tables and paragraphs. In addition to creating a user friendly interface to *L^AT_EX*, *R Markdown* also allows you to read in your data, to analyze it and to visualize it using **R** functions, and also to provide the documentation and commentary on the results of your project. Further, it allows for **R** results to be passed inline to the commentary of your results. You'll see more on this later.

Who should use it?

Anyone who needs to use data analysis, math, tables, a lot of figures, complex cross-references, or who just cares about the final appearance of their document should use *R Markdown*. Of particular use should be anyone in the sciences, but the user-friendly nature of *Markdown* and its ability to keep track of and easily include figures, automatically generate a table of contents, index, references, table of figures, etc. should make it of great benefit to nearly anyone writing a thesis project.

Chapter 1

EvoMining

EvoMining needs. Functional genomics uses genomes in the study of gene and 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. <http://rmarkdown.rstudio.com>.

1.1 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.

- 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
- BBH Best Bidirectional Hits with studied enzymes from Central Actinobacterial pathways were selected.
- By abundance
- By expansions on genomes

1.2 Natural Products DB

Natural products was improved from previous version

1.3 Reproducibility on EvoMining code

EvoMining Code was packaged on Docker ## 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 expansions. 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

1.4 Otras estrategias para los clusters Argon context Idea

Argon 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

1.5 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:

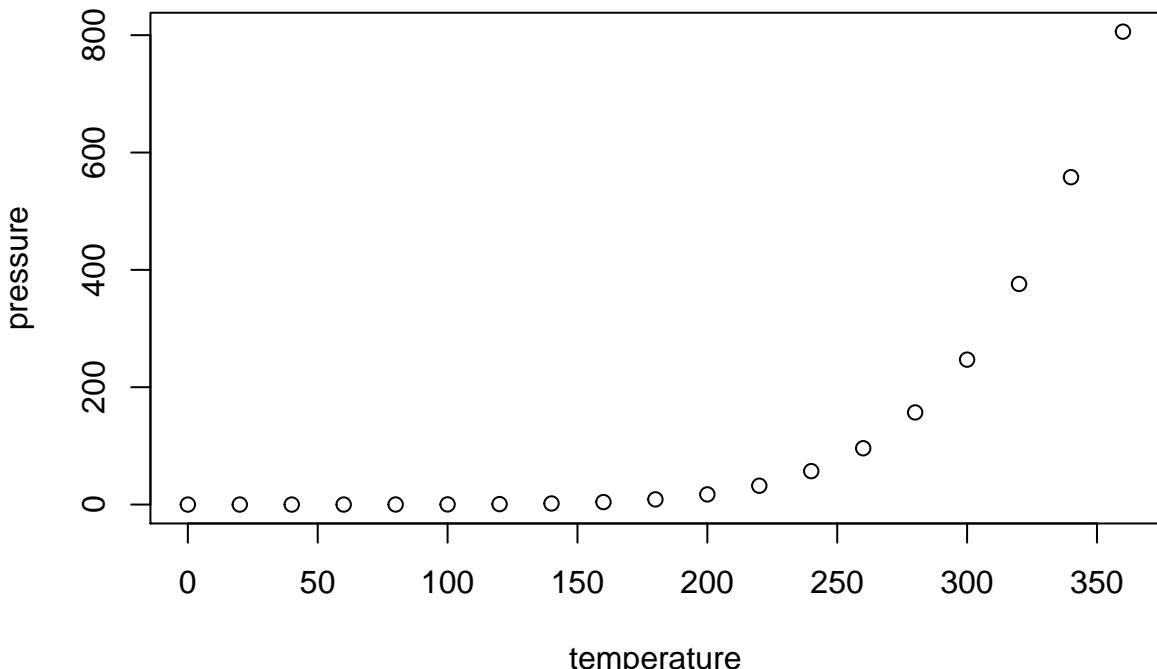
The standard deviation is less than 6.

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

As you see with `2π` 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.

1.6 CORASoN

You can also embed plots. For example, here is a way to use the base **R** graphics package to produce a plot using the built-in `pressure` dataset:



Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the **R** code that generated the plot. There are plenty of other ways to add chunk options. More information is available at <http://yihui.name/knitr/options/>.

Another useful chunk option is the setting of `cache = TRUE` as you see here. If document rendering becomes time consuming due to long computations or plots that are expensive to generate you can use knitr caching to improve performance. Later in this file, you'll see a way to reference plots created in **R** or external figures.

1.7 Loading and exploring data

Included in this template is a file called `flights.csv`. This file includes a subset of the larger dataset of information about all flights that departed from Seattle and Portland in 2014. More information about this dataset and its **R** package is available at <http://github.com/ismayc/pnwflights14>. This subset includes only Portland flights and only rows that were complete with no missing values. Merges were also done with the `airports` and `airlines` data sets in the `pnwflights14` package to get more descriptive airport and airline names.

We can load in this data set using the following command:

```
flights <- read.csv("data/flights.csv")
```

The data is now stored in the data frame called `flights` in **R**. To get a better feel for the variables included in this dataset we can use a variety of functions. Here we can see the dimensions (rows by columns) and also the names of the columns.

```
dim(flights)
```

```
[1] 52808    16
```

```
names(flights)
```

```
[1] "month"        "day"          "dep_time"      "dep_delay"  
[5] "arr_time"     "arr_delay"     "carrier"       "tailnum"  
[9] "flight"       "dest"         "air_time"      "distance"  
[13] "hour"         "minute"       "carrier_name" "dest_name"
```

Another good idea is to take a look at the dataset in table form. With this dataset having more than 50,000 rows, we won't explicitly show the results of the command here. I recommend you enter the command into the Console *after* you have run the **R** chunks above to load the data into **R**.

```
View(flights)
```

While not required, it is highly recommended you use the `dplyr` package to manipulate and summarize your data set as needed. It uses a syntax that is easy to understand using chaining operations. Below I've created a few examples of using `dplyr` to get information about the Portland flights in 2014. You will also see the use of the `ggplot2` package, which produces beautiful, high-quality academic visuals.

We begin by checking to ensure that needed packages are installed and then we load them into our current working environment:

```
# List of packages required for this analysis
pkg <- c("dplyr", "ggplot2", "knitr", "devtools")
# Check if packages are not installed and assign the
# names of the packages not installed to the variable new.pkg
new.pkg <- pkg[!(pkg %in% installed.packages())]
# If there are any packages in the list that aren't installed,
# install them
if (length(new.pkg))
  install.packages(new.pkg, repos = "http://cran.rstudio.com")
# Load packages
library(dplyr)
library(ggplot2)
library(knitr)
```

The example we show here does the following:

- Selects only the `carrier_name` and `arr_delay` from the `flights` dataset and then assigns this subset to a new variable called `flights2`.
- Using `flights2`, we determine the largest arrival delay for each of the carriers.

```
flights2 <- flights %>% dplyr::select(carrier_name, arr_delay)
max_delays <- flights2 %>% group_by(carrier_name) %>%
  summarize(max_arr_delay = max(arr_delay, na.rm = TRUE))
```

We next introduce a useful function in the `knitr` package for making nice tables in *R Markdown* called `kable`. It produces the `LATEX` code required to make the table and is much easier to use than manually entering values into a table by copying and pasting values into Excel or `LATEX`. This again goes to show how nice reproducible documents can be! There is no need to copy-and-paste values to create a table. (Note the use of `results = "asis"` here which will produce the table instead of the code to create the table. You'll learn more about the `\label` later.) The `caption.short` argument is used to include a shorter version of the title to appear in the List of Tables at the beginning of the document.

```
kable(max_delays, col.names = c("Airline", "Max Arrival Delay"),
      caption = "Maximum Delays by Airline \\label{tab:max_delay}",
      caption.short = "Max Delays by Airline")
```

Table 1.1: Maximum Delays by Airline

Airline	Max Arrival Delay
Alaska Airlines Inc.	338
American Airlines Inc.	1539
Delta Air Lines Inc.	651
Frontier Airlines Inc.	575
Hawaiian Airlines Inc.	407
JetBlue Airways	273
SkyWest Airlines Inc.	421
Southwest Airlines Co.	694
United Air Lines Inc.	472
US Airways Inc.	347
Virgin America	366

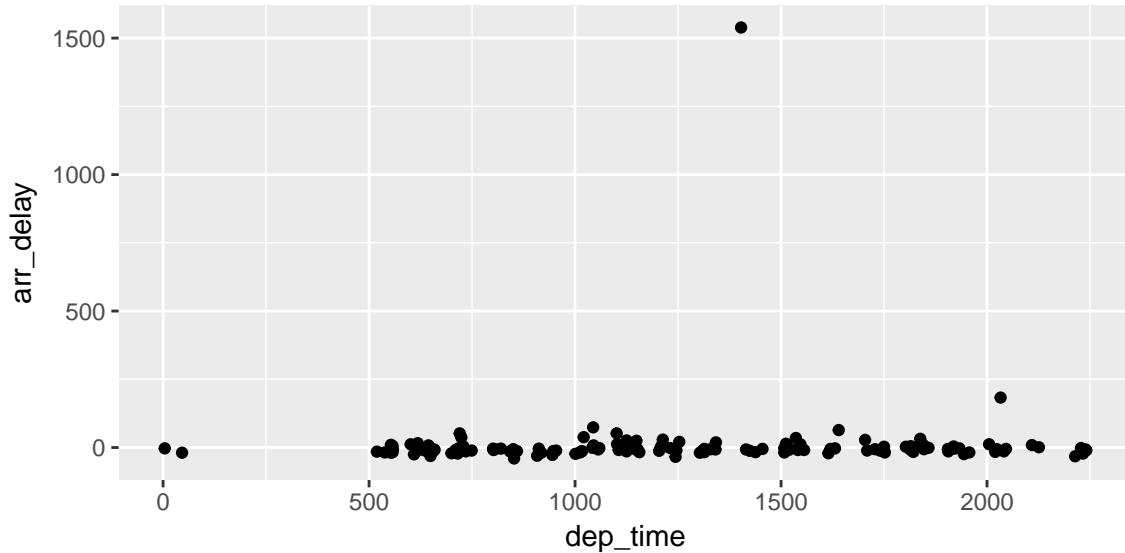
We can further look into the properties of the largest value here for American Airlines Inc. To do so, we can isolate the row corresponding to the arrival delay of 1539 minutes for American in our original `flights` dataset.

```
flights %>% dplyr::filter(arr_delay == 1539,
                           carrier_name == "American Airlines Inc.") %>%
  dplyr::select(-c(month, day, carrier, dest_name, hour,
                  minute, carrier_name, arr_delay))
```

	dep_time	dep_delay	arr_time	tailnum	flight	dest	air_time	distance
1	1403	1553	1934	N595AA	1568	DFW	182	1616

We see that the flight occurred on March 3rd and departed a little after 2 PM on its way to Dallas/Fort Worth. Lastly, we show how we can visualize the arrival delay of all departing flights from Portland on March 3rd against time of departure.

```
flights %>% dplyr::filter(month == 3, day == 3) %>%
  ggplot(aes(x = dep_time, y = arr_delay)) +
  geom_point()
```



1.8 Additional resources

- *Markdown Cheatsheet* - <https://github.com/adam-p/markdown-here/wiki/Markdown-Cheatsheet>
- *R Markdown Reference Guide* - <https://www.rstudio.com/wp-content/uploads/2015/03/rmarkdown-reference.pdf>
- Introduction to dplyr - <https://cran.rstudio.com/web/packages/dplyr/vignettes/introduction.html>
- ggplot2 Documentation - <http://docs.ggplot2.org/current/>

Chapter 2

PriA Family

- Julian simulation
- Who has TrpF

2.1 Math

\TeX is the best way to typeset mathematics. Donald Knuth designed \TeX when he got frustrated at how long it was taking the typesetters to finish his book, which contained a lot of mathematics. One nice feature of *R Markdown* is its ability to read \LaTeX code directly.

If you are doing a thesis that will involve lots of math, you will want to read the following section which has been commented out. If you're not going to use math, skip over or delete this next commented section.

2.2 Chemistry 101: Symbols

Chemical formulas will look best if they are not italicized. Get around math mode's automatic italicizing in \LaTeX by using the argument $\$\\mathrm{formula\ here}\\$$, with your formula inside the curly brackets. (Notice the use of the backticks here which enclose text that acts as code.)

So, $\text{Fe}_2^{2+}\text{Cr}_2\text{O}_4$ is written $\$\\mathrm{Fe_2^{2+}Cr_2O_4}\\$$.
Exponent or Superscript: O^-
Subscript: CH_4

To stack numbers or letters as in Fe_2^{2+} , the subscript is defined first, and then the superscript is defined.

Angstrom: \AA
Bullet: $\text{CuCl} \bullet 7\text{H}_2\text{O}$
Double Dagger: \ddagger
Delta: Δ
Reaction Arrows: \longrightarrow or $\xrightarrow{\text{solution}}$
Resonance Arrows: \leftrightarrow

Reversible Reaction Arrows: \rightleftharpoons or $\xrightleftharpoons[solution]{}$ (the latter requires the `chemarr` L^AT_EX package which is automatically loaded in this template)

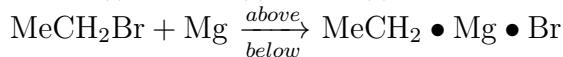
2.2.1 Typesetting reactions

You may wish to put your reaction in a figure environment, which means that L^AT_EX will place the reaction where it fits and you can have a figure caption. You'll see further description of this `R label` function in . (Note the use of the double backslash here as well as the `echo = FALSE` which hides the code from the output.)



Figure 2.1: Combustion of glucose

2.2.2 Other examples of reactions



2.3 Physics

Many of the symbols you will need can be found on the math page <http://web.reed.edu/cis/help/latex/math.html> and the Comprehensive L^AT_EX Symbol Guide (<http://mirror.utexas.edu/ctan/info/symbols/comprehensive/symbols-letter.pdf>).

2.4 Biology

You will probably find the resources at <http://www.lecb.ncifcrf.gov/~toms/latex.html> helpful, particularly the links to bsts for various journals. You may also be interested in TeXShade for nucleotide typesetting (<http://homepages.uni-tuebingen.de/beitz/txe.html>). Be sure to read the proceeding chapter on graphics and tables.

Chapter 3

Archaea EvoMining Results

Archaea kingdom was discovered by Woese in 1977 (Woese & Fox, 1977)
Actinobacteria has proved that natural products can be found on expansions from central metabolic pathways
Archaea is interesting because as we will see has a phylum that contains expansions (Central and pangenomic)
on Actinobacteria previously known natural products can be detected by software like antiSMASH and clusterfinder, nevertheless on Archaea a whole kingdom, this products are not found. Does Archaea not produce NP's or traditionally known method are nos suitable for finding them? Archea has operons??
Archaea has introns?

3.1 Tables

Table 3.1: Families on Archaeabacteria

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

3.1.1 Expansions BoxPlot by metabolic family

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

Here is a reference to the expansion boxplot: Figure 3.1.

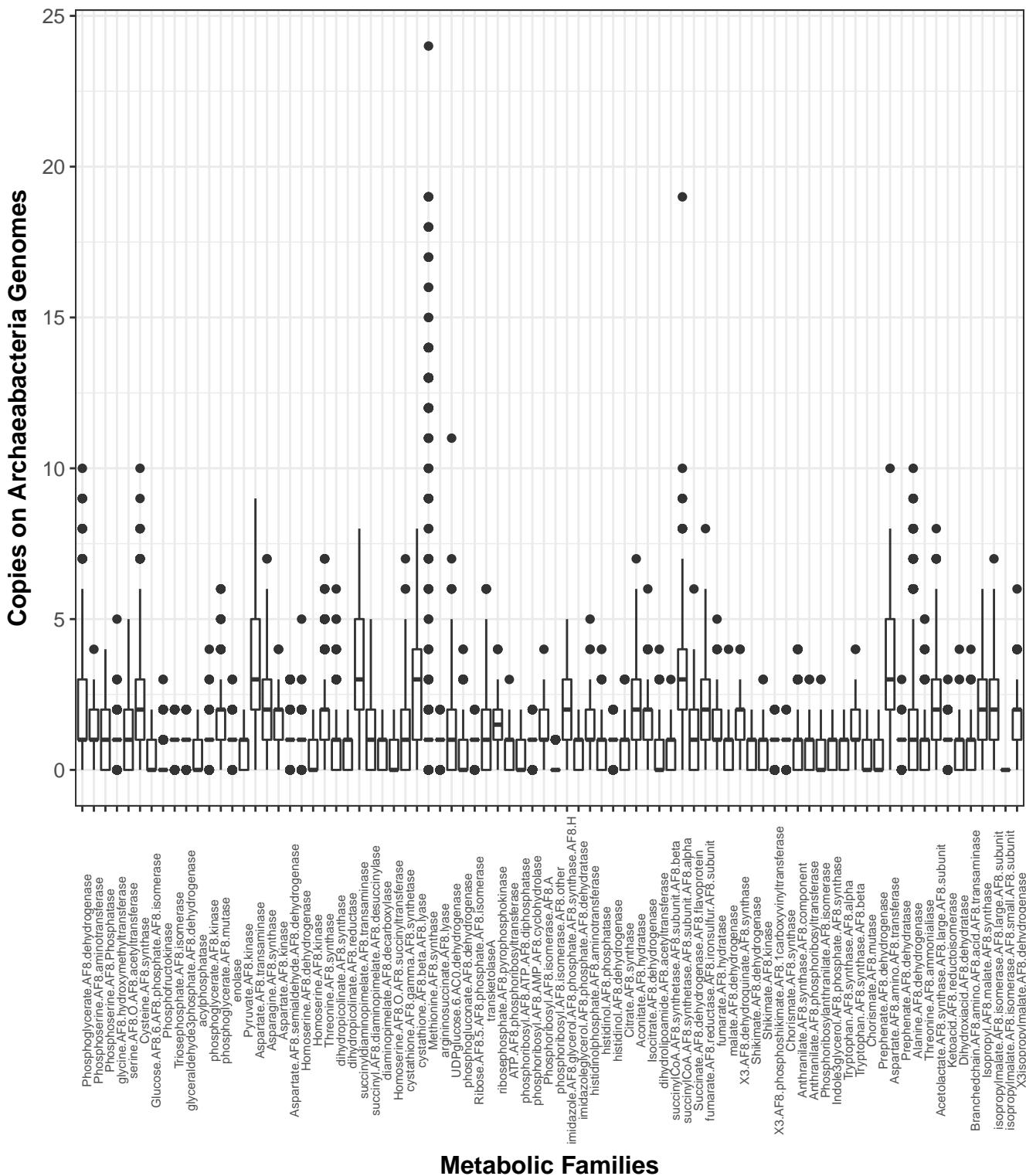


Figure 3.1: Expansions Boxplot

3.2 Central pathway expansions

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

Warning: Ignoring unknown aesthetics: width

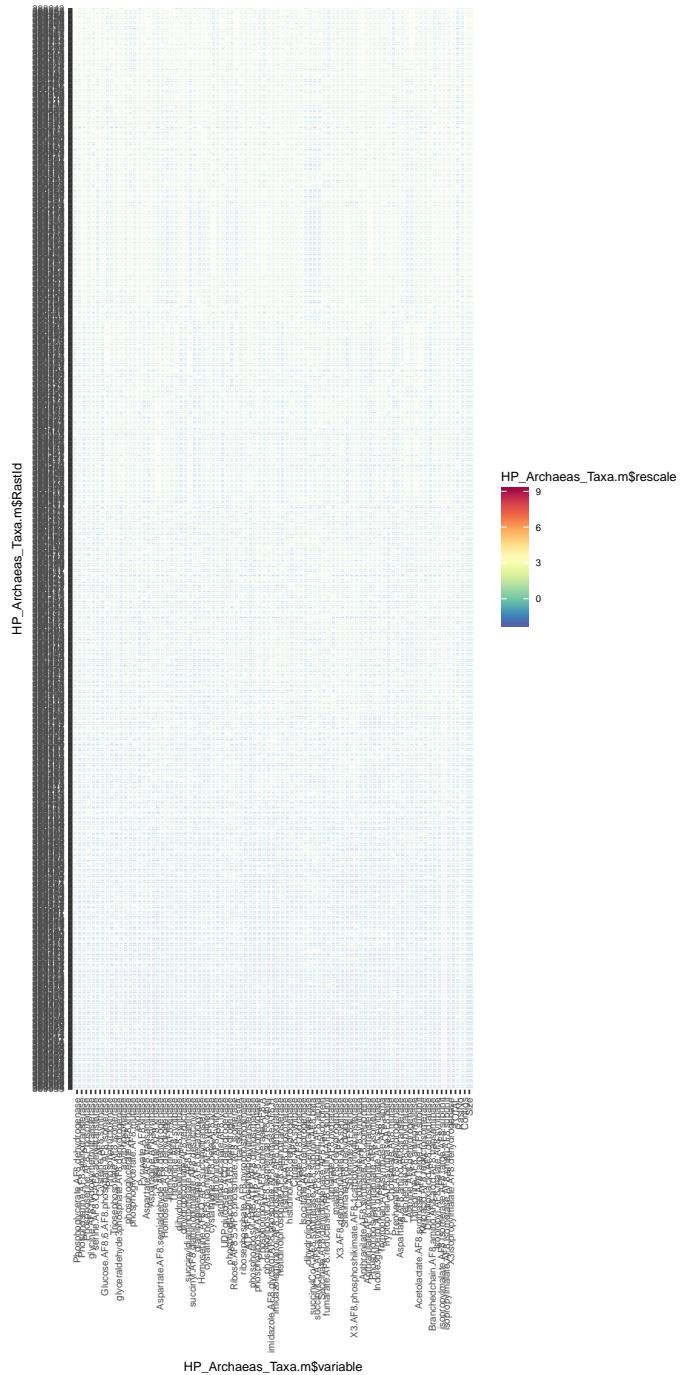


Figure 3.2: Archaeas Heatplot

Here is a reference to the HeatPlot: Figure 3.2.

3.3 Genome Size correlations

3.3.1 Correlation between genome size and AntiSMASH products

Genome size vs Total antismash cluster coloured by order

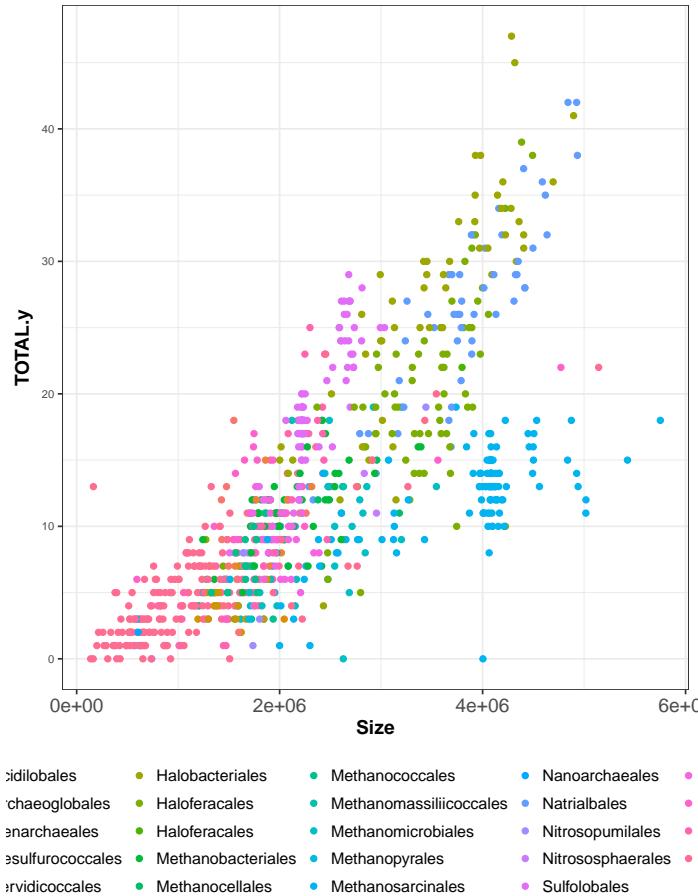


Figure 3.3: Correlation between Archaea genome size and antismash Natural products detection colored by Order

Here is a reference to Genome size vs Total antismash cluster: Figure 3.3.

Genome size vs Total antismash cluster detected splitted by order

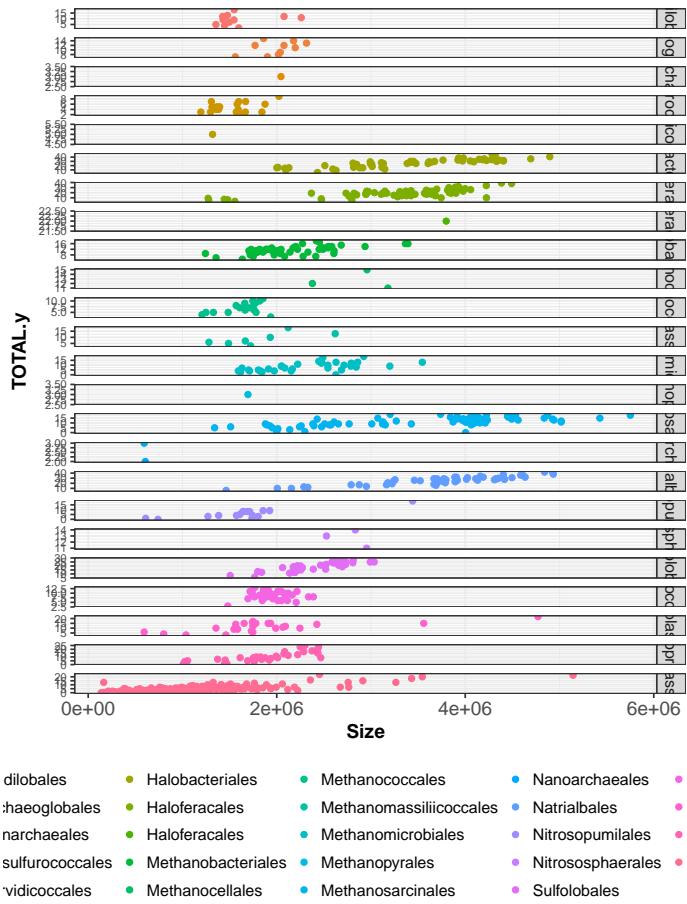


Figure 3.4: Correlation between Archaeas genome size and antismash Natural products detection grided by Order

Here is a reference to Correlation between genome size and antismash Natural products detection grided by Order plot: Figure 3.4.

3.3.2 Correlation between genome size and Central pathway expansions

Genome size vs Total central pathway expansion coloured by order

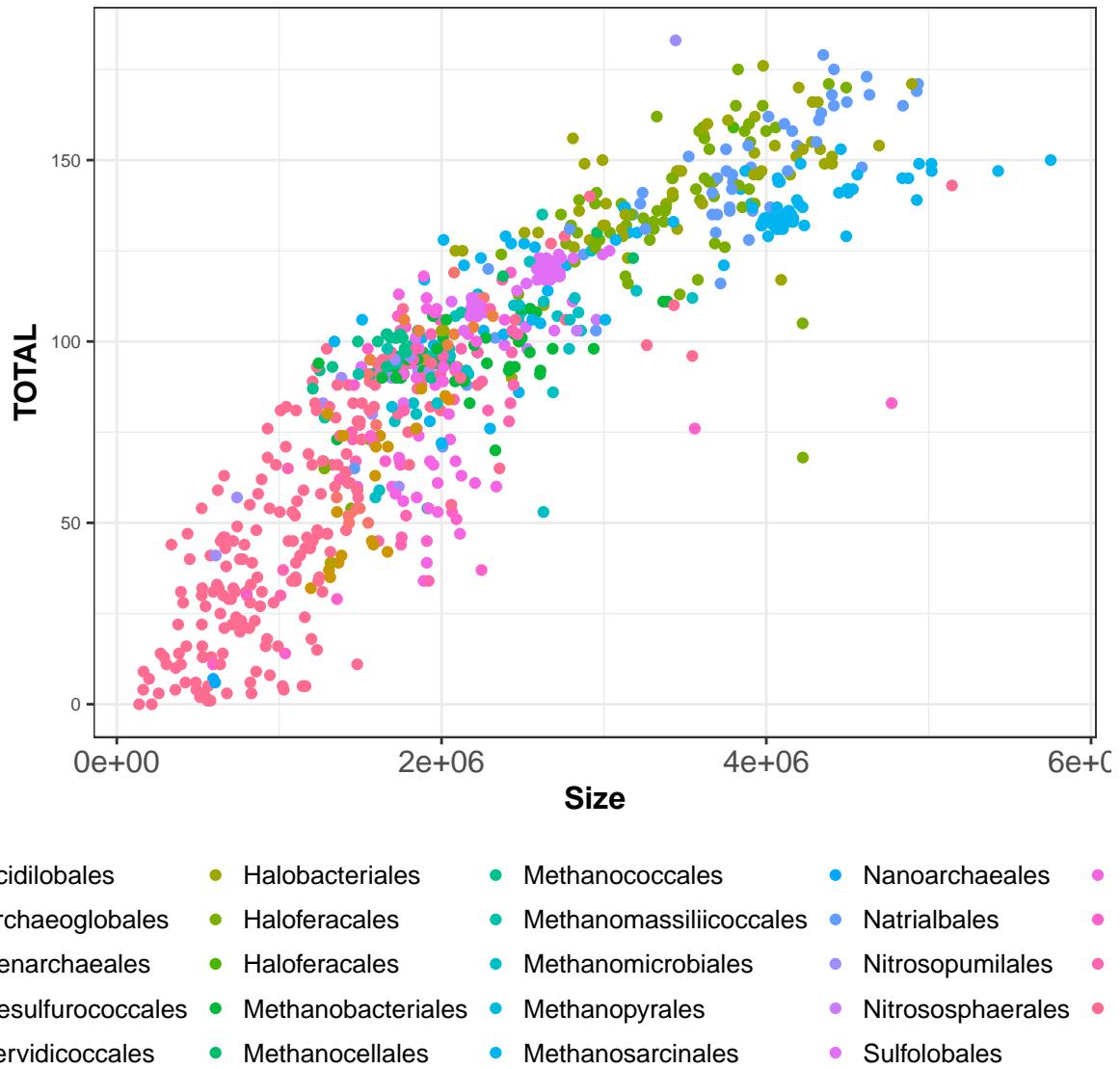


Figure 3.5: Correlation between Archaea genome size and central pathway expansions

Here is a reference to the size vs Total central pathway expansion plot: Figure 3.5.

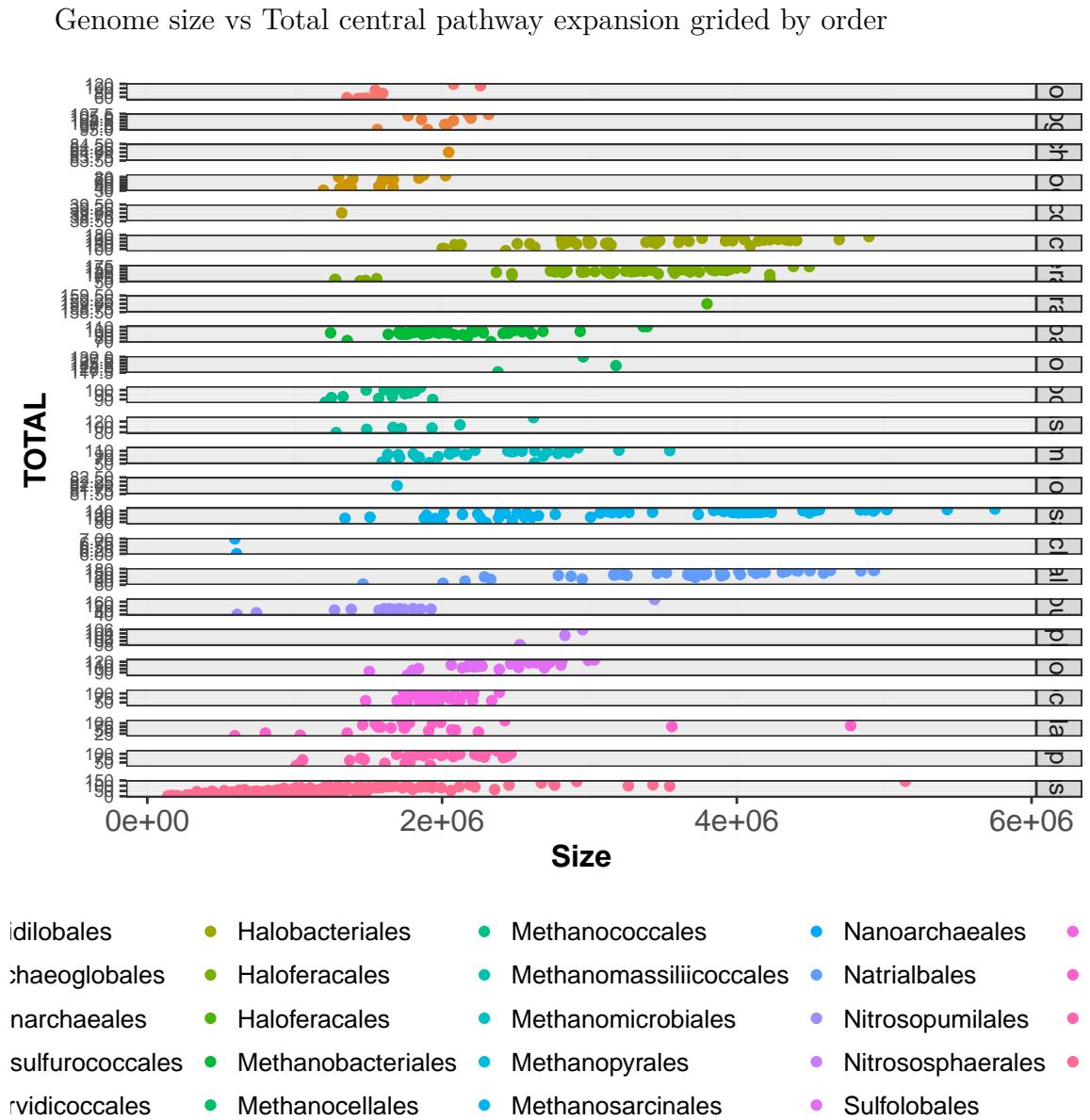


Figure 3.6: Correlation between Archaea genome size and central pathway expansions grided by order

Here is a reference to the Genome size vs Total central pathway expansion grided by order plot: Figure 3.6.

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

Also I want to add form given by taxonomical order.

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

Warning: Removed 64823 rows containing missing values (geom_point).

Genome size vs Total central pathway expansion coloured by metabolic Family

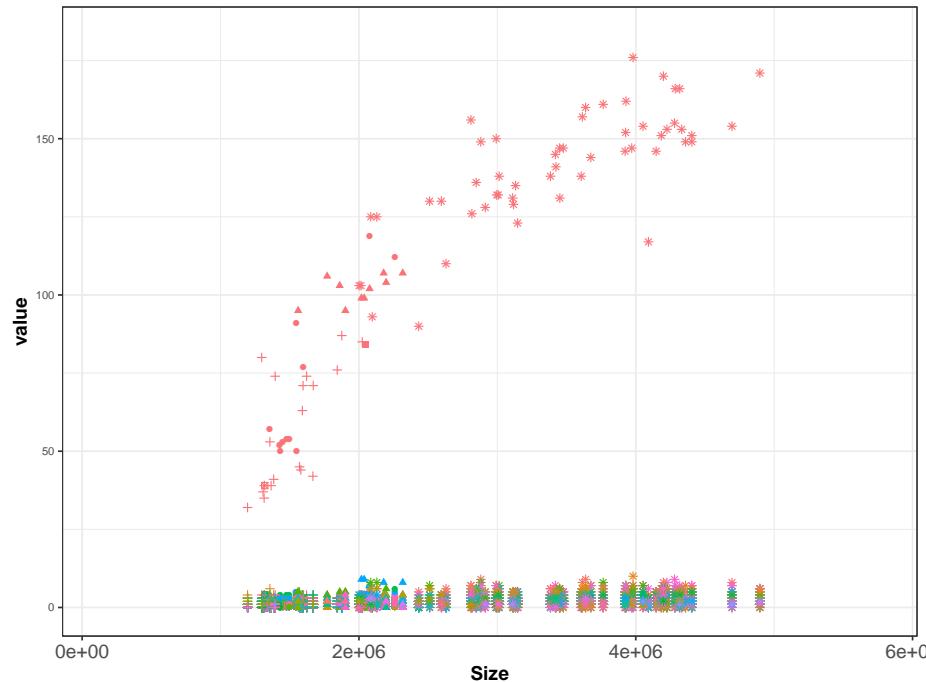


Figure 3.7: Correlation between Archaea's Genome size vs Total central pathway expansion coloured by metabolic Family

Here is a reference to the Genome size vs Total central pathway expansion coloured by metabolic Family plot: Figure 3.7.

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: ??.

3.4 Natural products

3.4.1 Natural products recruitments from EvoMining heat-plot

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

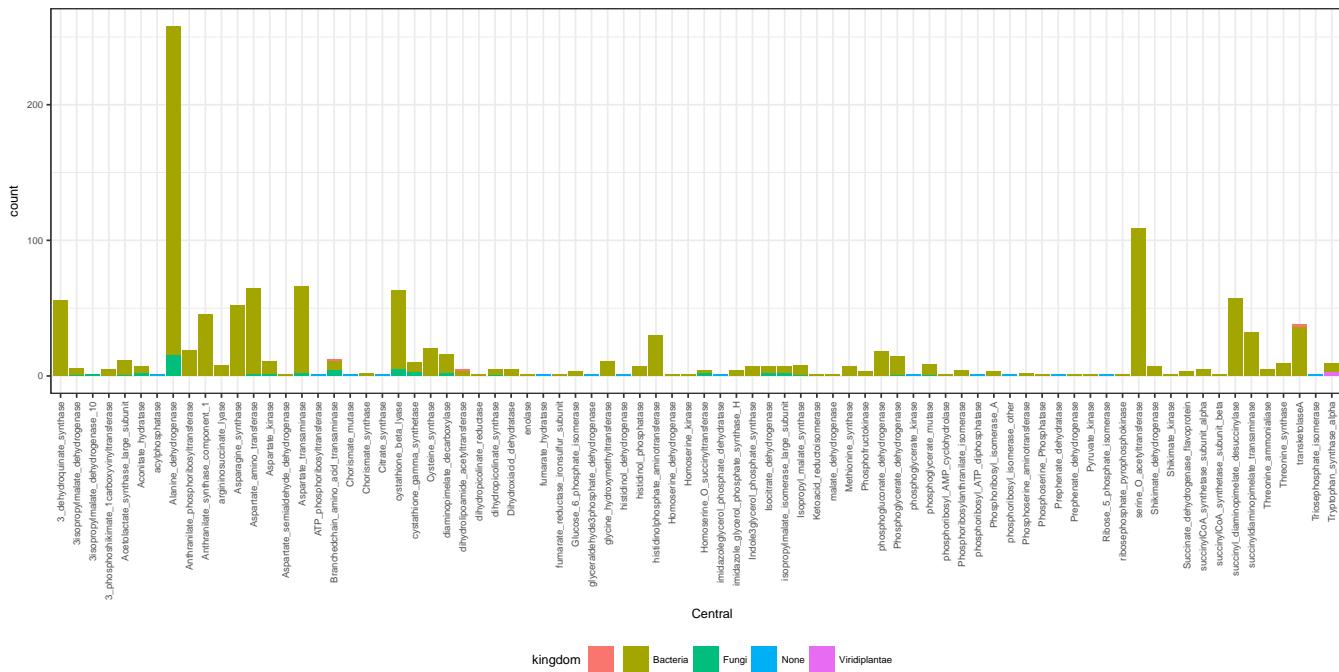


Figure 3.8: Archaeas Recruitmens on central families coloured by kingdom

Here is a reference to Recruitments after central pathways expansions colourd by Kingdom plot: Figure 3.8.

Recruitments after central pathways expansions coloured by taxonomy

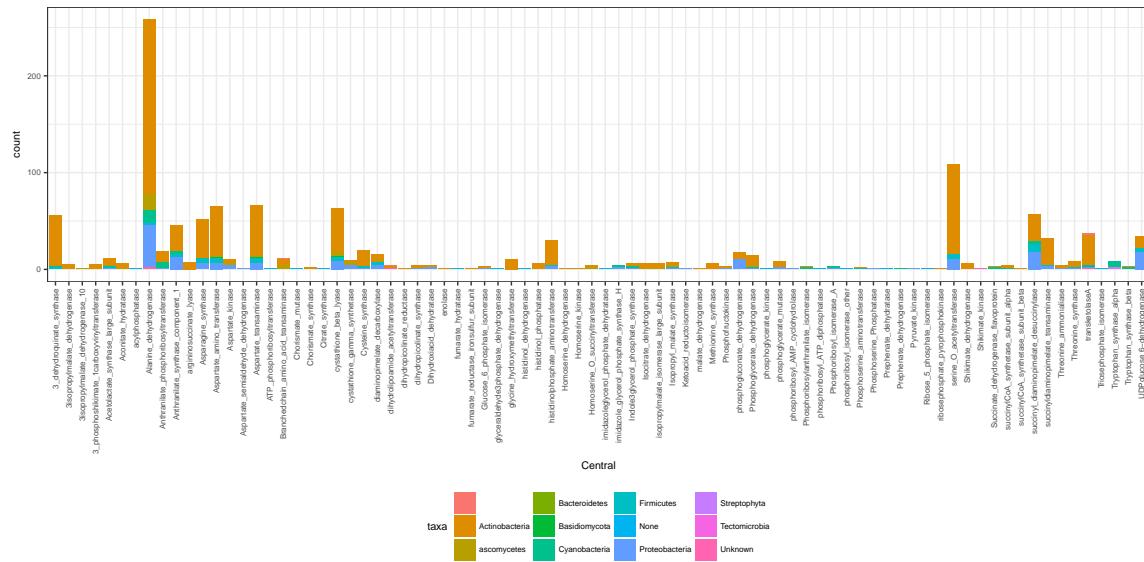


Figure 3.9: Archaeas Recruitmens on central families coloured by taxonomy

Here is a reference to Recruitments after central pathways expansions coloured by taxa plot: Figure 3.9.

3.5 Archaeas AntiSMASH

Taxonomical diversity on Archaeasbacteria Data

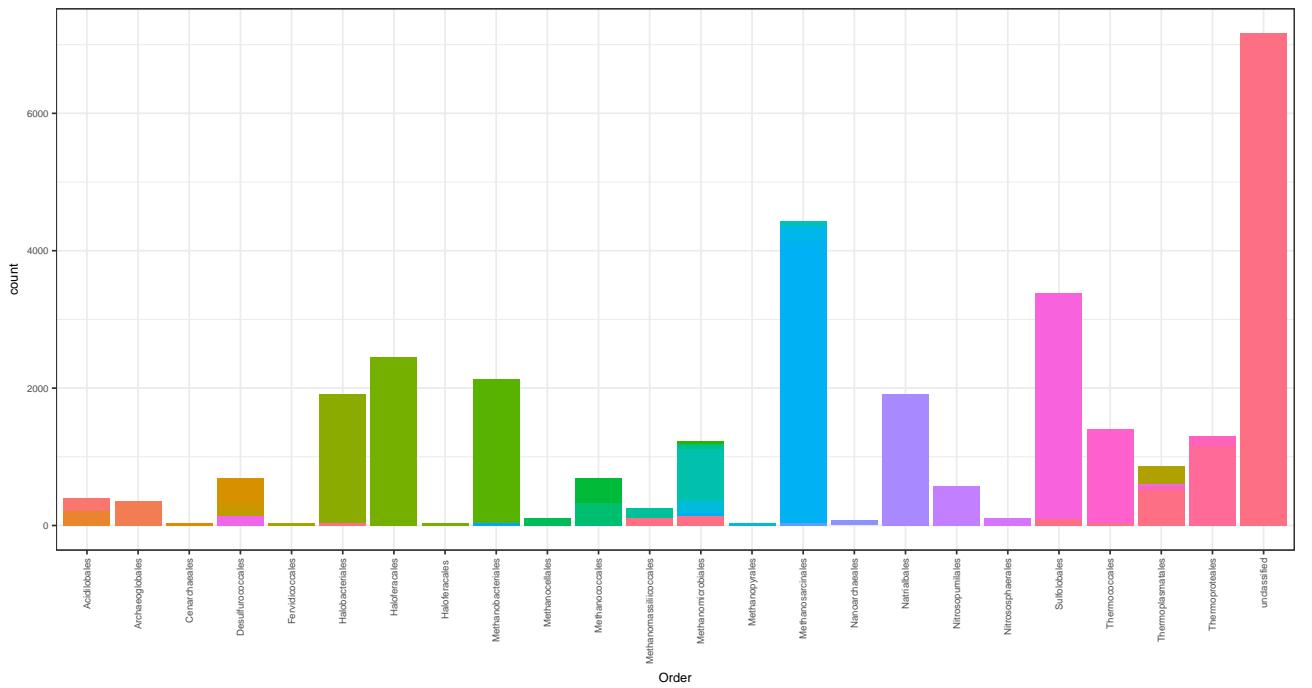


Figure 3.10: Archaeas Diversity

Here is a reference to Recruitments after central pathways expansions coloured by taxa plot: Figure 3.10.

Smash diversity

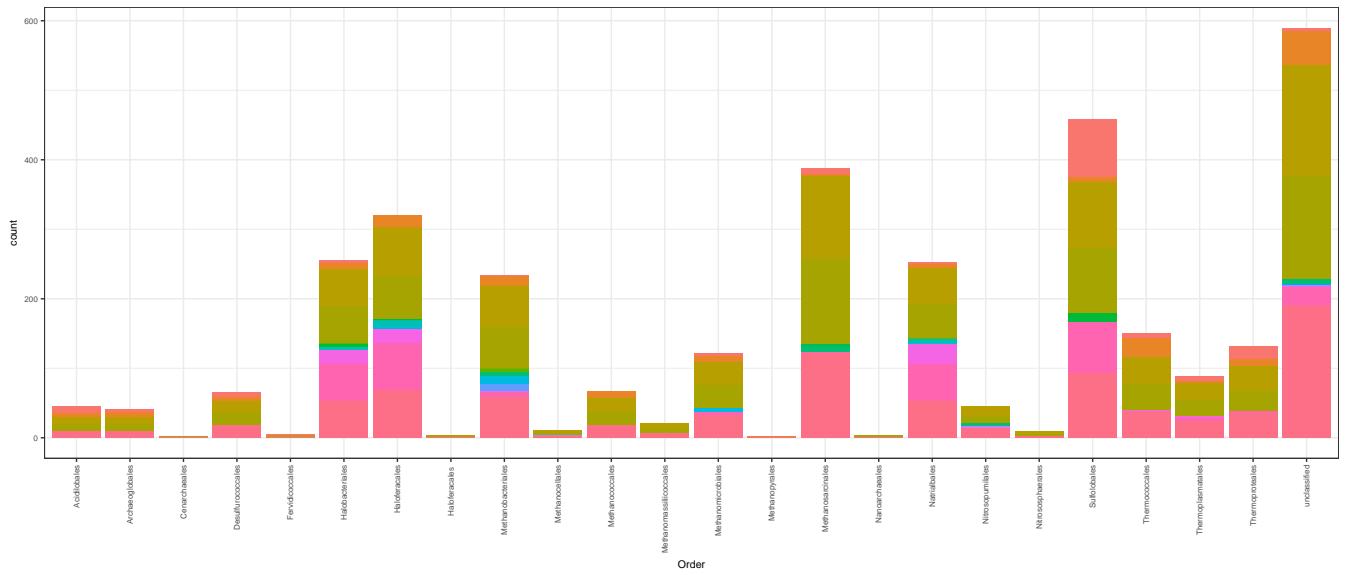


Figure 3.11: Archaea Smash Taxonomical Diversity

Here is a reference to Recruitments after central pathways expansions coloured by taxa plot: Figure 3.11.

3.5.1 AntisMASH vs Central Expansions

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

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

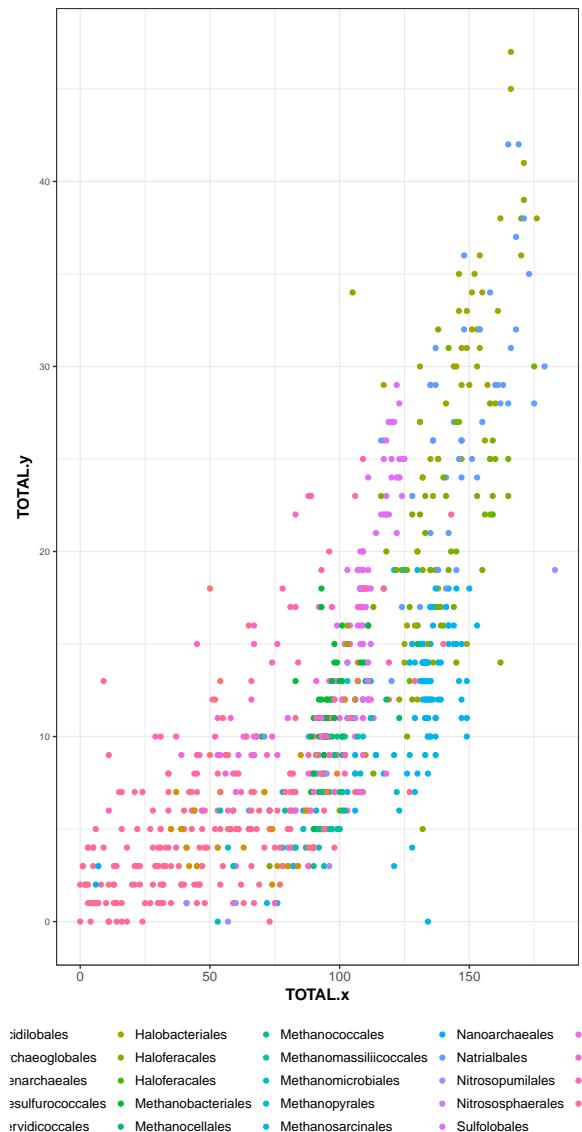


Figure 3.12: Correlation between Archaeas central pathway expansions and antismash Natural products detection

Here is a reference to the expansions vs antimash NP's clusters plot: Figure 3.12.

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

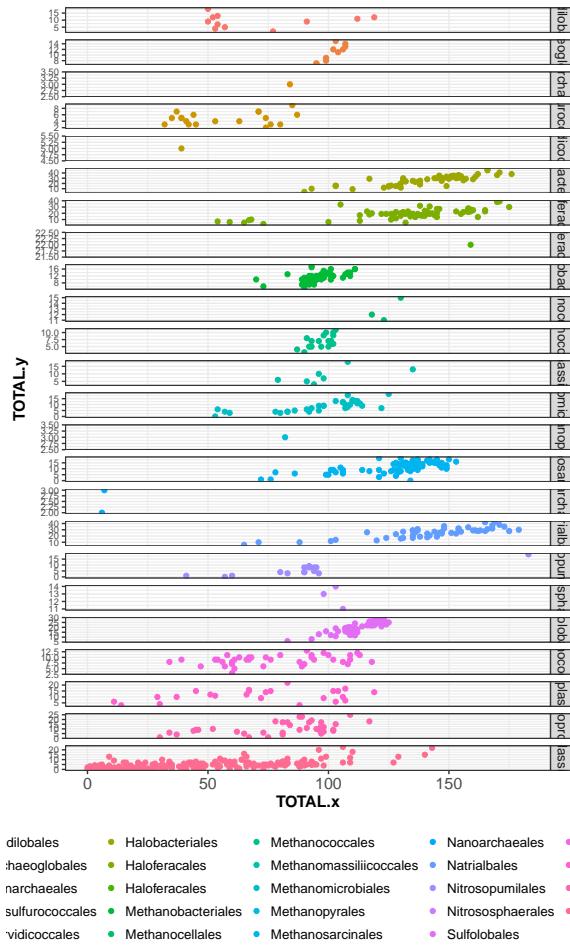


Figure 3.13: Correlation between Archaeas central pathway expnasions and antismash NP's clusters splitted by order

Here is a reference to the expansions vs antismash NP's clusters splitted by order plot Figure 3.13.

AntisMAsh vs Expansions by taxonomic Family Natural products colured by family

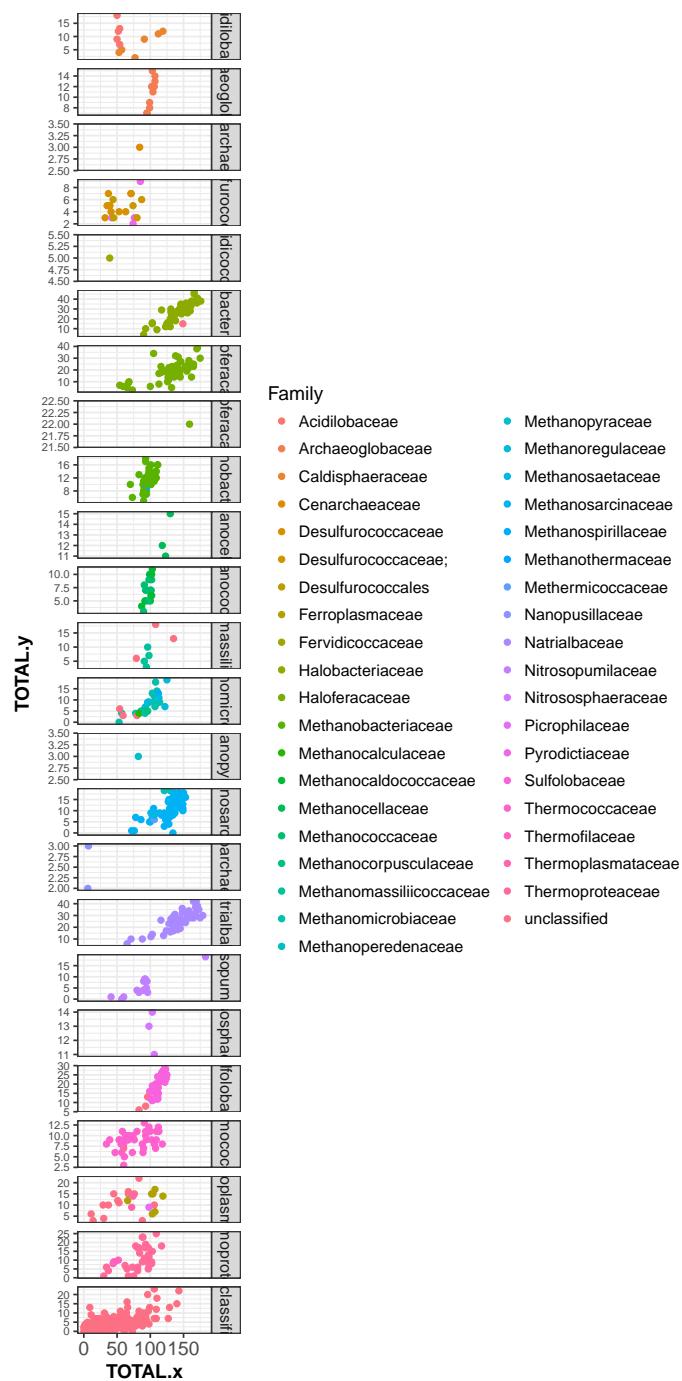


Figure 3.14: Archaeas Natural products by family

Here is a reference to the Natural products coloured by family plot Figure 3.14.

3.6 Selected trees from EvoMining

Phosphoribosyl_isomerase_3 family

Figure from EvoMining

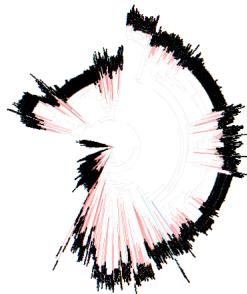


Figure 3.15: Phosphoribosyl isomerase A EvoMiningtree

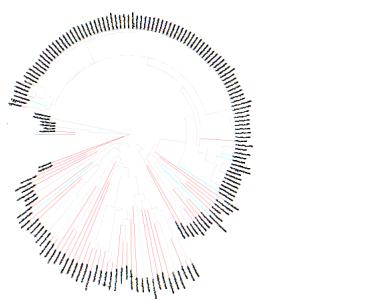


Figure 3.16: Phosphoribosyl isomerase other EvoMiningtree

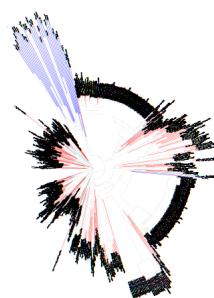


Figure 3.17: Phosphoribosyl anthranilate isomerase EvoMiningtree

3.7 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.

3.8 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/zotermiddlebury/>.

R Markdown uses *pandoc* (<http://pandoc.org/>) to build its bibliographies. One nice caveat of this is that you won't have to do a second compile to load in references as standard L^AT_EX 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: (Molina & Borkovec, 1994). This Molina1994 entry appears in a file called **thesis.bib** in the **bib** folder. This bibliography database file was created by a program called BibTeX. You can call this file something else if you like (look at the YAML header in the main .Rmd file) and, by default, is placed in the **bib** folder.

For more information about BibTeX and bibliographies, see our CUS site (<http://web.reed.edu/cis/help/latex/index.html>)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.

¹footnote text

²Reed College (2007)

- The cite key (a citation’s label) needs to be unique from the other entries.
- When you have more than one author or editor, you need to separate each author’s name by the word “and” e.g. Author = {Noble, Sam and Youngberg, Jessica},.
- Bibliographies made using BibTeX (whether manually or using a manager) accept L^AT_EX 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 citation³ 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.”

3.9 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.

³Noble (2002)

Chapter 4

Actinobacteria EvoMining Results

Actinobacteria is an ancient phylum {Referencia de luis}

4.1 Tables

Table 4.1: Correlation of Inheritance Factors for Parents and Child

Factors	Correlation between Parents & Child
GenomeDB	1245
Families	65

4.1.1 Expansions BoxPlot by metabolic family

```
label(path = "chapter4/expansion_plotActinos.pdf", caption = "Expansions Boxplot", label
```

Here is a reference to the expansion boxplot: Figure 4.1.

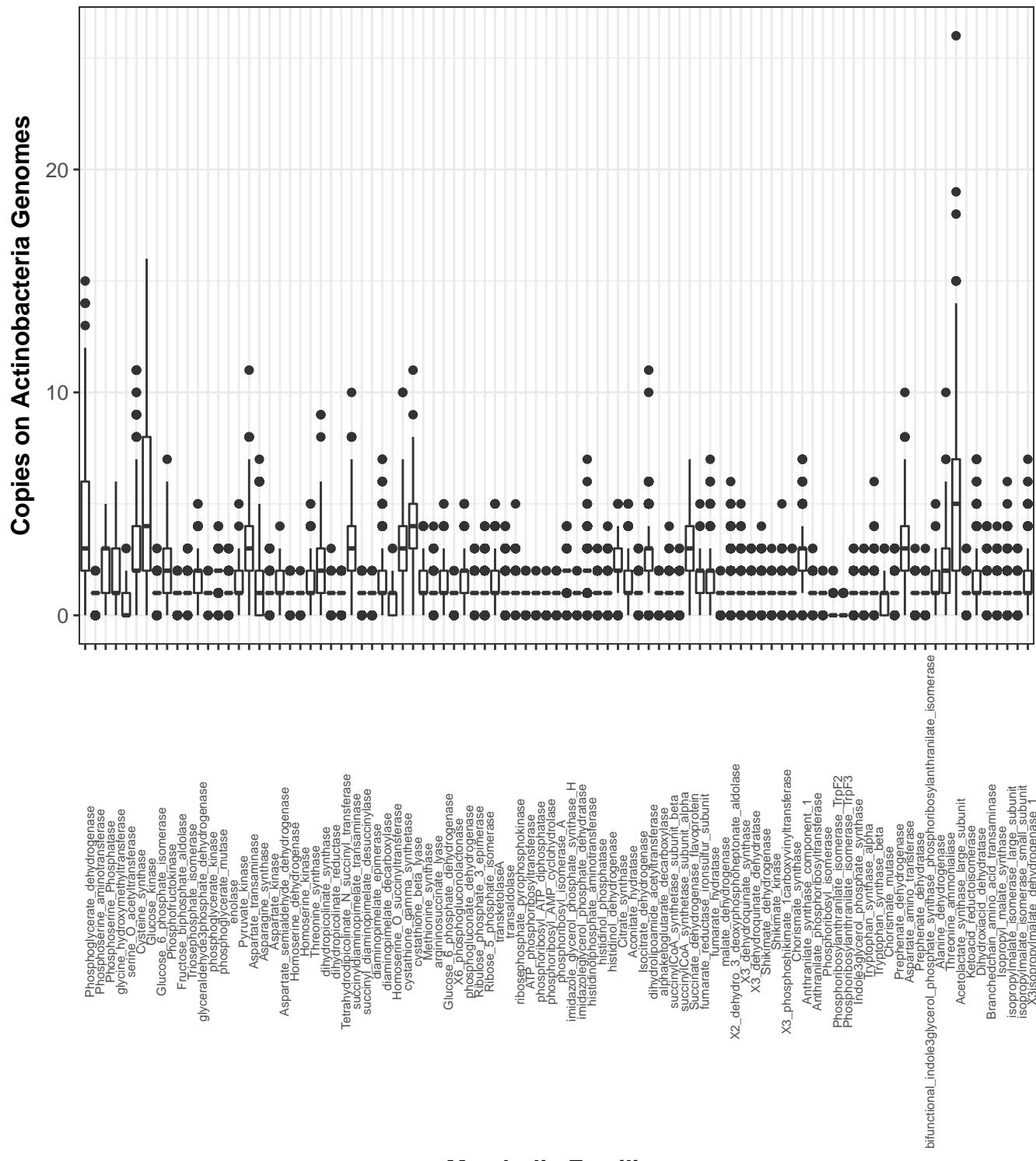
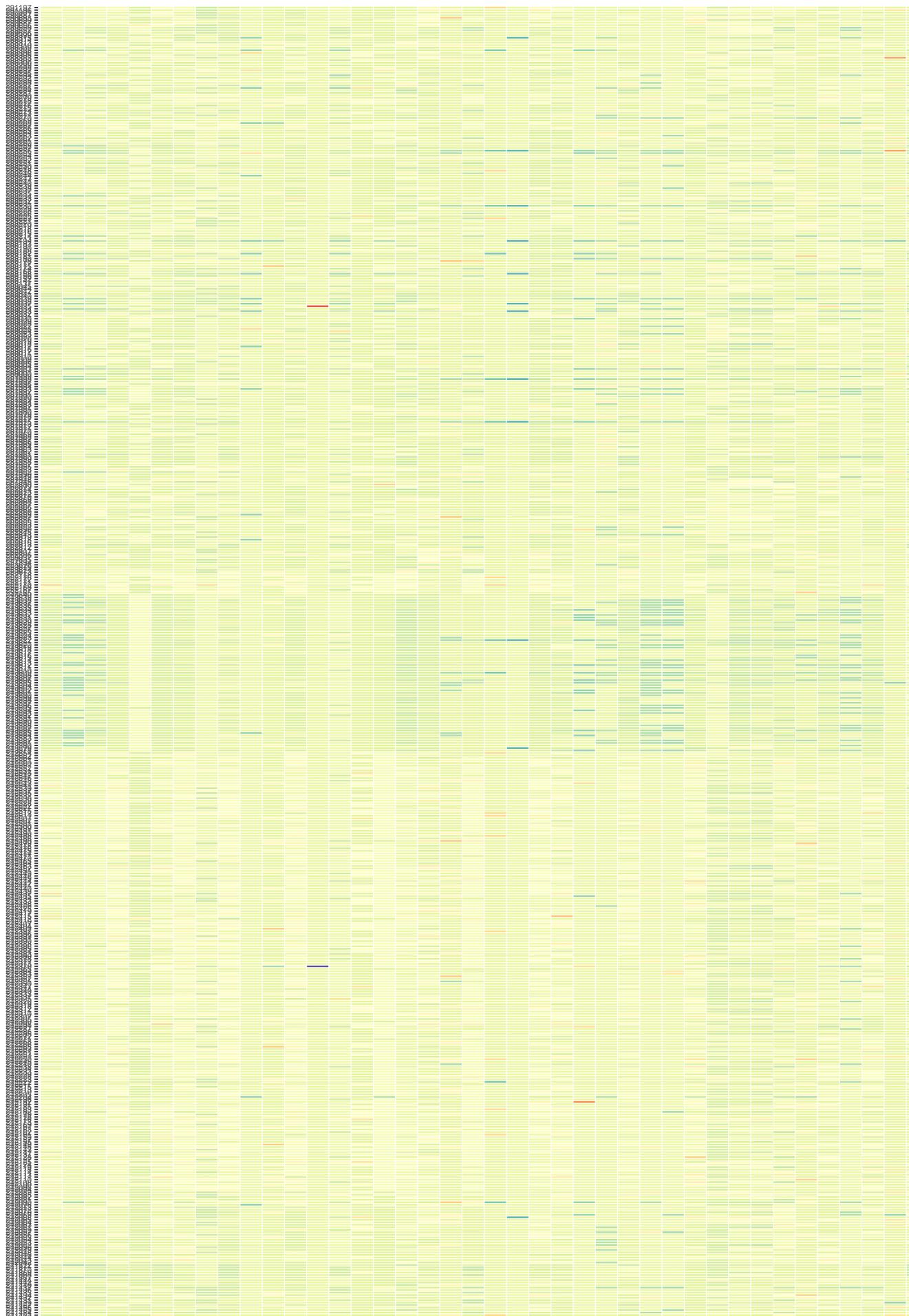


Figure 4.1: Expansions Boxplot

4.2 Central pathway expansions

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

Here is a reference to the HeatPlot: Figure 4.2.



PPP pathway expansions restricted to *Streptomycetaceae* family HeatPlot: Figure 4.2.

Here is a reference to the HeatPlot: Figure 4.3.

288310		
288306		
288302		
288308		
288289		
288280		
288278		
288277		
288268		
288267		
288266		
288261		
288254		
288233		
288211		
288218		
288215		
288188		
288178		
288162		
288032		
288019		
288012		
287996		
287984		
287979		
287965		
287955		
287953		
287950		
287948		
282855		
252178		
252176		
252172		
252171		
252170		
252168		
252167		
252165		
242654		
242652		
242564		
242561		
242560		
242557		
242552		
242551		
242548		
242547		
242546		
242545		
242543		
242541		
242539		
242537		
242535		
242532		
242530		
242529		
242528		
242522		
242521		
242515		
242513		
242511		
242507		
242502		
242501		
242500		
242499		
242491		
242490		
242488		
242486		
242480		
242479		
242478		
242476		
242475		
242474		
242471		
242470		
242465		
242464		
242463		
242452		
242451		
242449		
242448		
242445		
242444		
242442		
242441		
242440		
242438		
242437		
242435		
242434		
242433		
242428		
242426		
242425		
242419		
242417		
242415		
242412		
242410		
242407		
242404		
242402		
242397		
242396		
242393		
242391		
242390		
242388		
242386		
242385		
242384		
242383		
242380		
242378		
242377		
242373		
242370		
242365		
242364		
242363		
242357		
242352		
242351		
242350		
242347		
242344		
242340		
242333		
242331		
242325		
242320		
242319		
242318		
242312		
242311		
242310		
242307		
242305		
242300		
242298		
242291		
242287		
242286		
242285		
242272		
242271		
242268		
242266		
242265		
242263		
242261		
242253		
242251		
242250		
242245		
242240		

4.3 Genome Size correlations

4.3.1 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

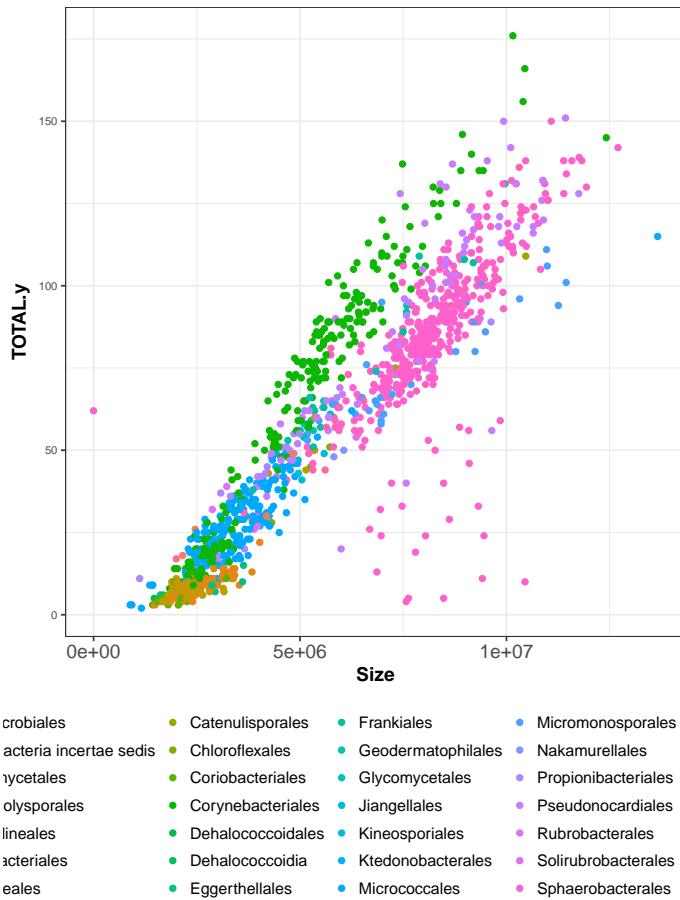


Figure 4.4: Correlation between Actinos genome size and antismash Natural products detection colored by Order

Here is a reference to Genome size vs Total antismash cluster: Figure 4.4.

Genome size vs Total antismash cluster detected splitted by order

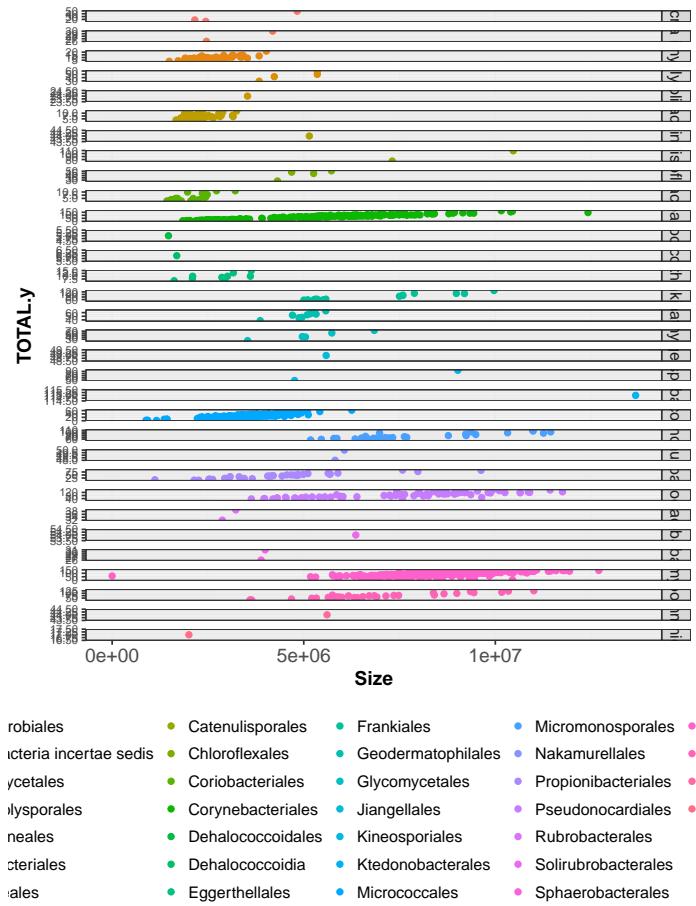


Figure 4.5: Correlation between Actinos genome size and antismash Natural products detection grided by Order

Here is a reference to Correlation between genome size and antismash Natural products detection grided by Order plot: Figure 4.5.

4.3.2 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

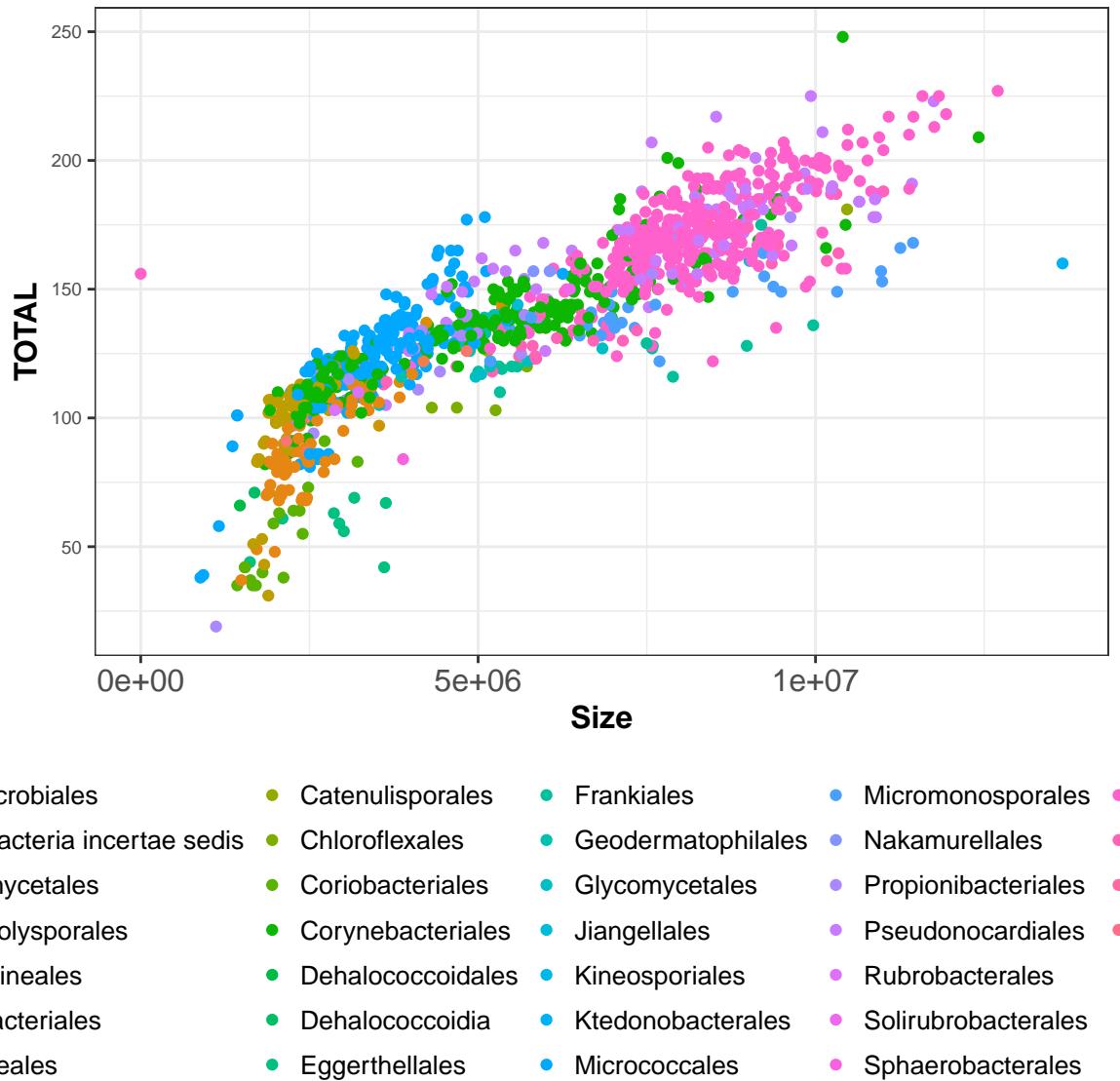


Figure 4.6: Correlation between Actinos genome size and central pathway expansions

Here is a reference to the size vs Total central pathway expansion plot: Figure 4.6.

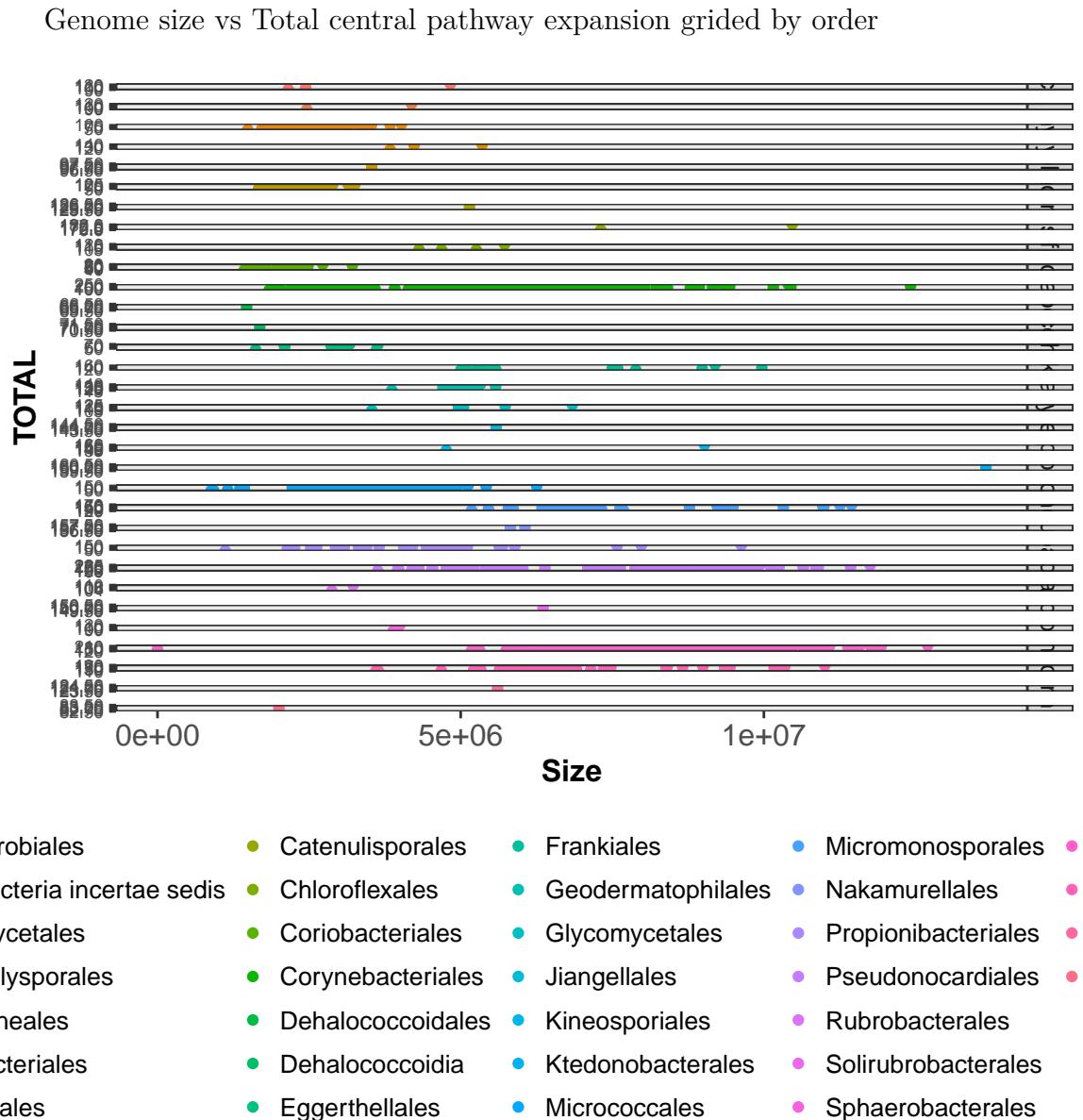


Figure 4.7: Correlation between Actinos genome size and central pathway expansions grided by order

Here is a reference to the Genome size vs Total central pathway expansion grided by order plot: Figure 4.7.

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

Also I want to add form given by taxonomical order.

Warning: The shape palette can deal with a maximum of 6 discrete values because more than 6 becomes difficult to discriminate; you have 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

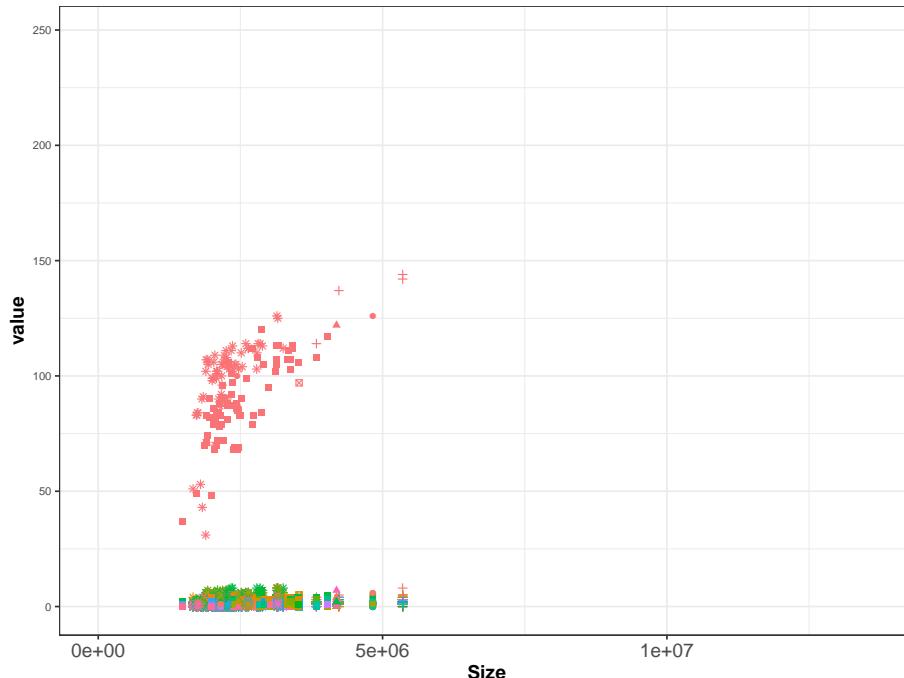


Figure 4.8: Correlation between Actinos Genome size vs Total central pathway expansion coloured by metabolic Family

Here is a reference to the Genome size vs Total central pathway expansion coloured by metabolic Family plot: Figure 4.8.

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: ??.

4.4 Natural products

4.4.1 Natural products recruitments from EvoMining heat-plot

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

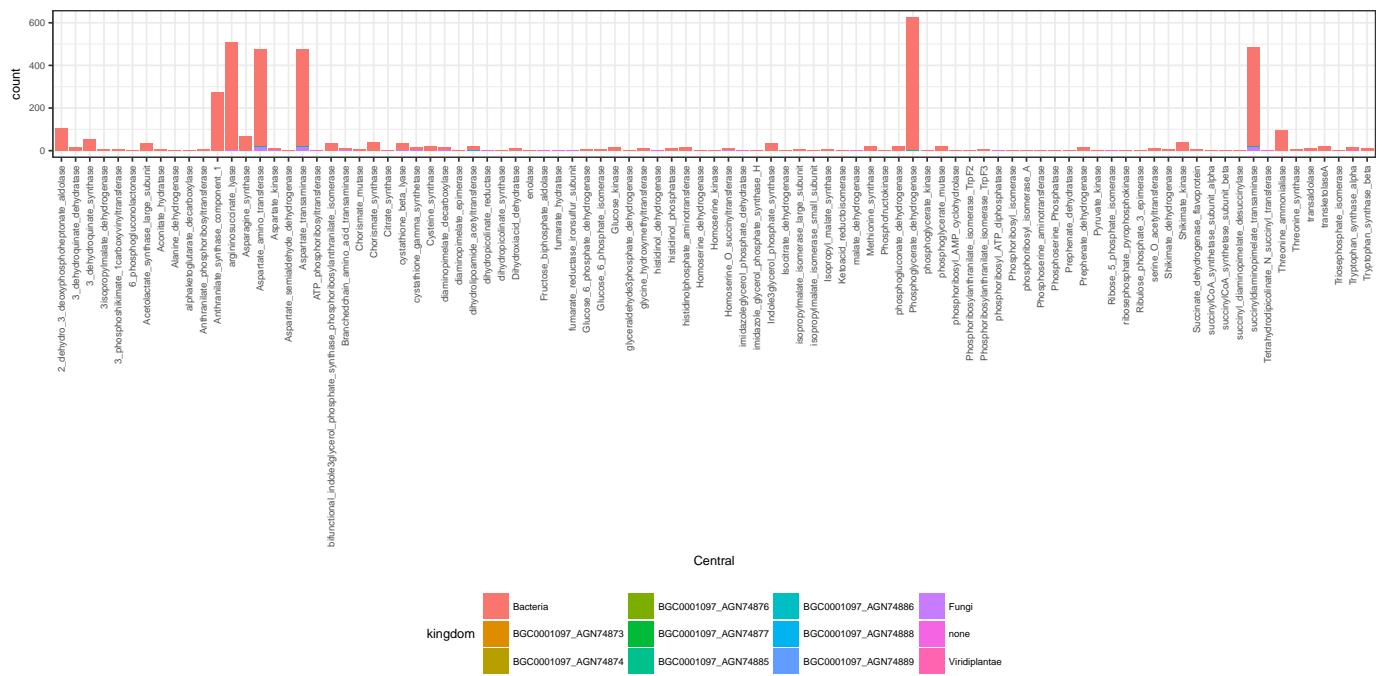


Figure 4.9: Actinos Recruitmens on central families coloured by kingdom

Here is a reference to Recruitments after central pathways expansions coloured by Kingdom plot: Figure 4.9.

Recruitments after central pathways expansions colourd by taxonomy

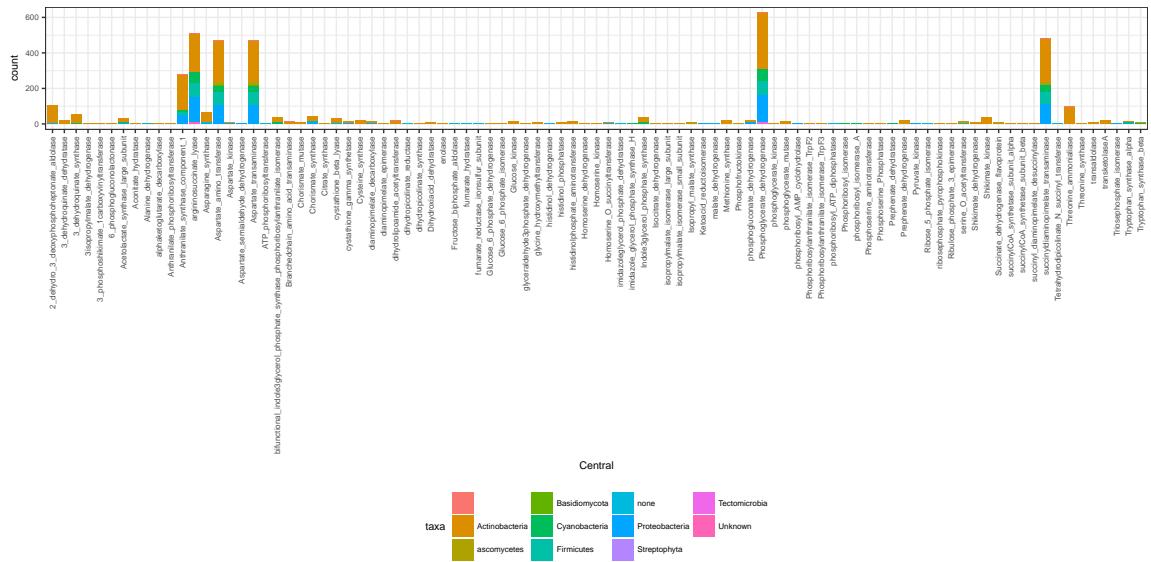


Figure 4.10: Actinos Recruitmens on central families coloured by taxonomy

Here is a reference to Recruitments after central pathways expansions coloured by taxa plot: Figure 4.10.

4.5 Actinos AntiSMASH

Taxonomical diversity on Actinosbacteria Data

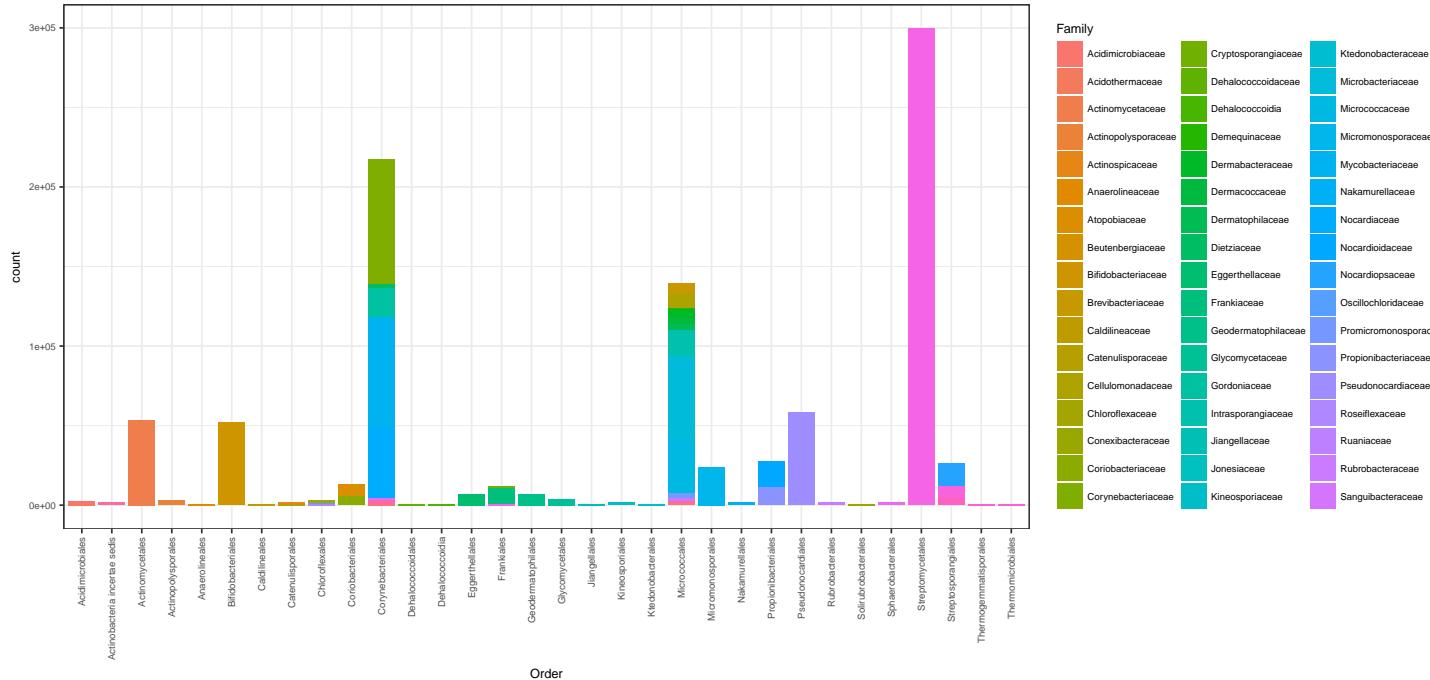


Figure 4.11: Actinos Diversity

Here is a reference to Recruitments after central pathways expansions coloured by taxa plot: Figure 4.11.

Smash diversity

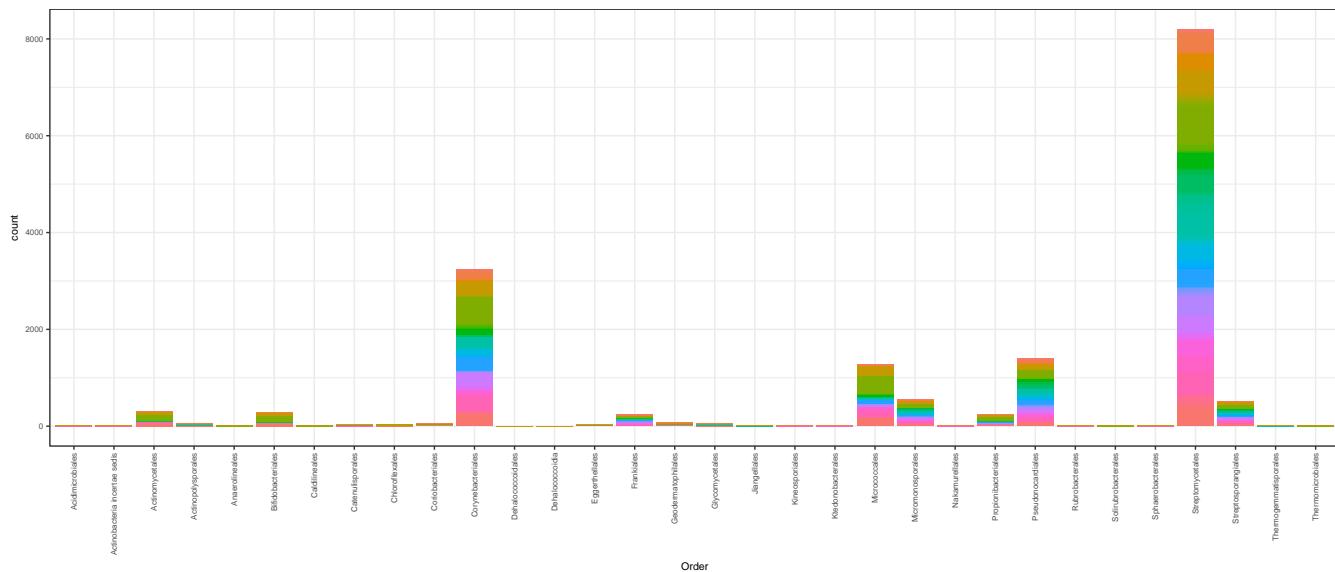


Figure 4.12: Actinos Smash Taxonomical Diversity

Here is a reference to Recruitments after central pathways expansions coloured by taxa plot: Figure 4.12.

4.5.1 AntisMASH vs Central Expansions

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

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

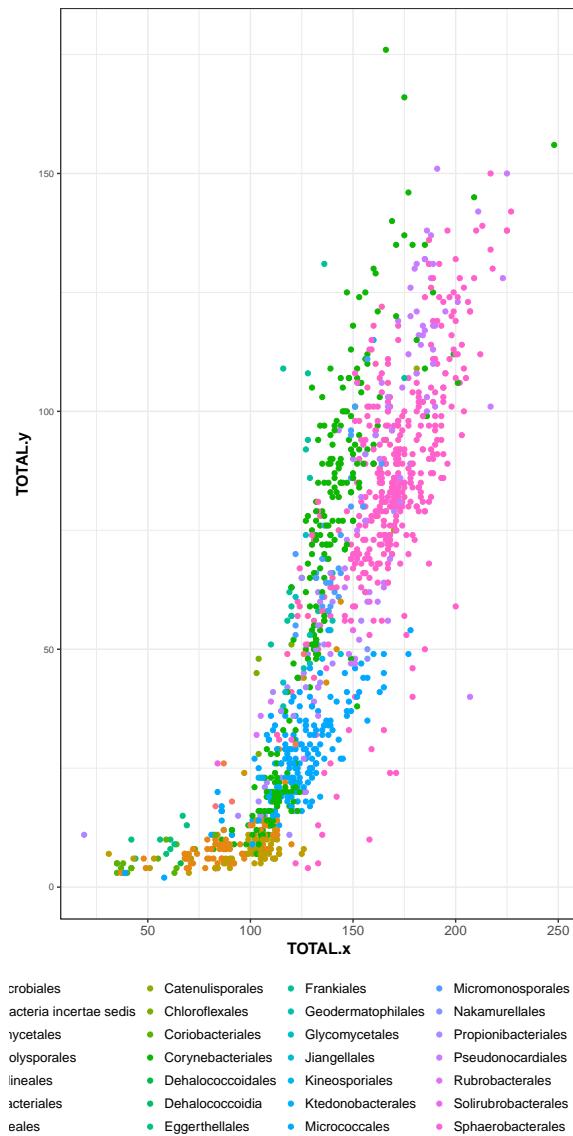


Figure 4.13: Correlation between Actinos central pathway expansions and antisimash Natural products detection

Here is a reference to the expansions vs antisimash NP's clusters plot: Figure 4.13.

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

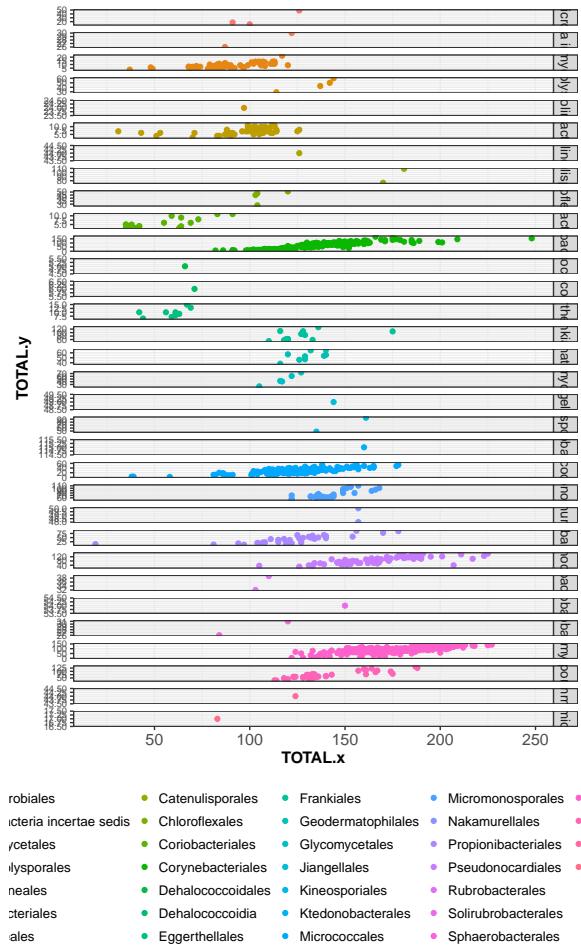


Figure 4.14: Correlation between Actinos central pathway expansions and antismash Natural products detection

Here is a reference to the expansions vs antismash NP's clusters splitted by order plot Figure 4.14.

AntisMAsh vs Expansions by taxonomic Family
 Natural products colured by family



Figure 4.15: Actinos Natural products by family

Here is a reference to the Natural products colured by family plot Figure 4.15.

4.6 Selected trees from EvoMining

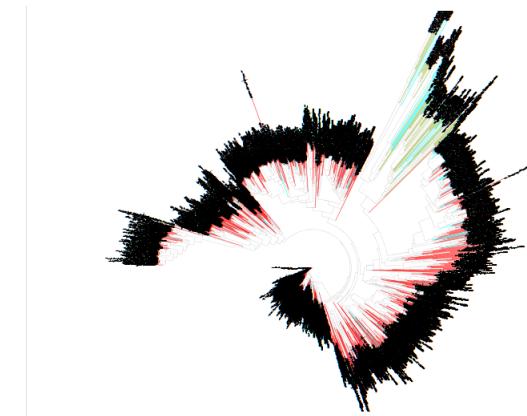


Figure 4.16: Enolase EvoMiningtree

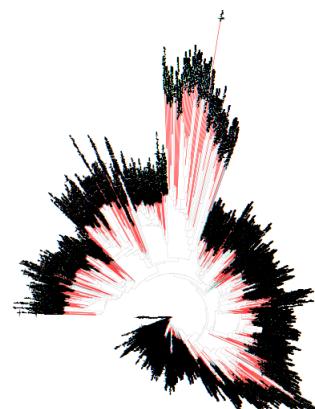


Figure 4.17: Phosphoribosyl isomerase EvoMiningtree

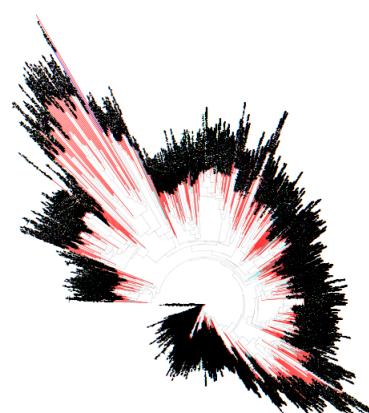


Figure 4.18: Phosphoribosyl isomerase A EvoMiningtree

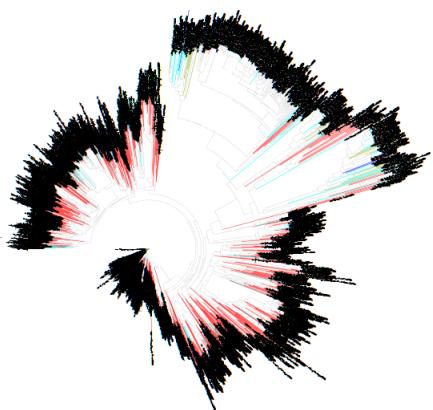


Figure 4.19: phosphoshikimate carboxyvinyltransferase EvoMiningtree

Chapter 5

Cyanobacteria EvoMining Results

Cyanobacteria phylum {Referencia}

5.1 Tables

Table 5.1: Families on Cyanobacteria

Factors	Correlation between Parents & Child
GenomeDB	1245
Families	65

5.1.1 Expansions BoxPlot by metabolic family

```
label(path = "chapter5/expansion_plotCyanos.pdf", caption = "Expansions Boxplot",label =
```

Here is a reference to the expansion boxplot: Figure 5.1.

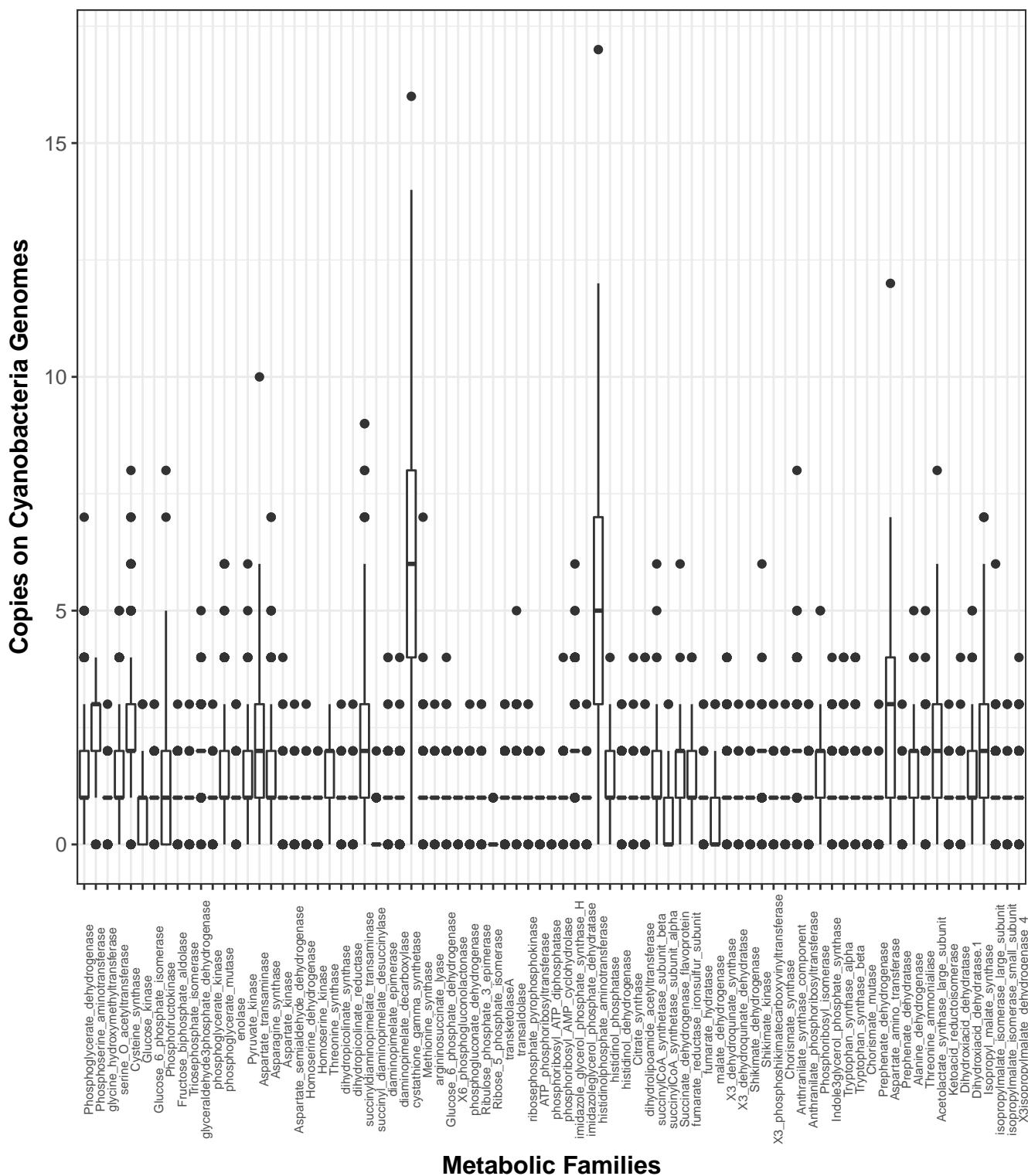


Figure 5.1: Expansions Boxplot

5.2 Central pathway expansions

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

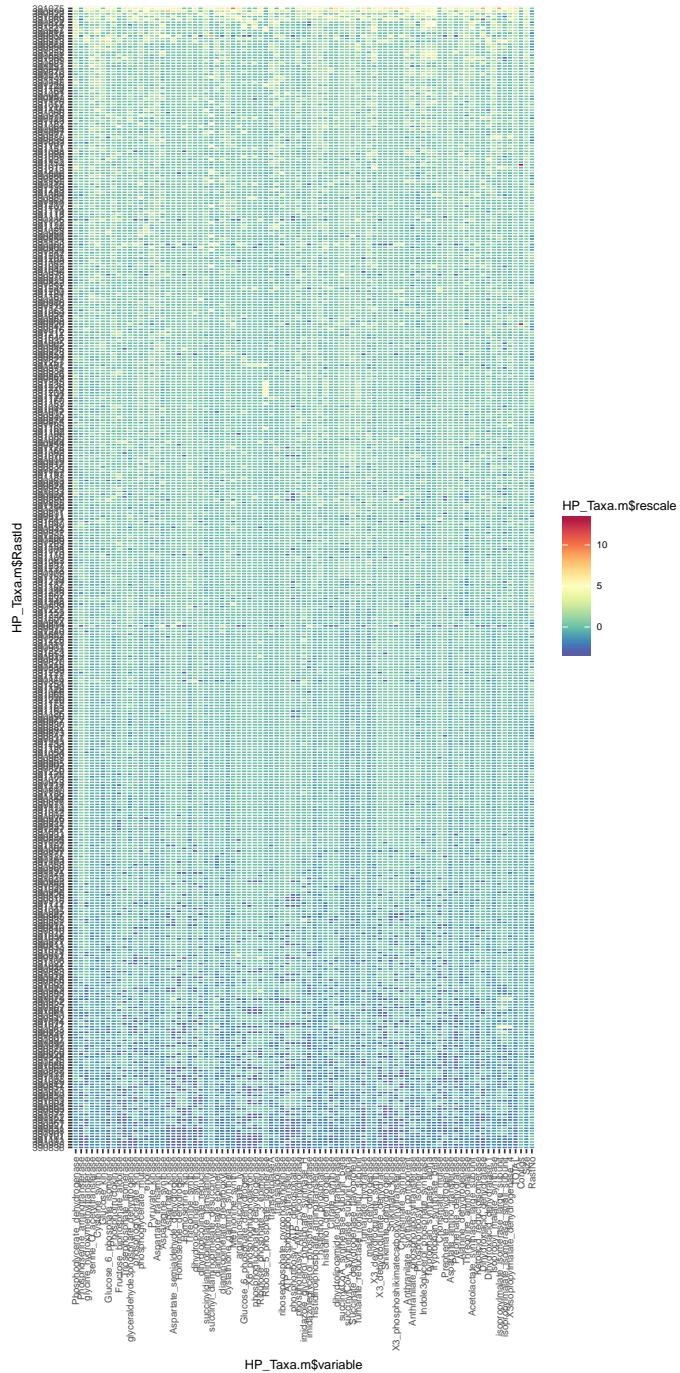


Figure 5.2: Cyanobacterial Heatplot

Here is a reference to the HeatPlot: Figure 5.2.

5.3 Genome Size correlations

5.3.1 Correlation between genome size and AntiSMASH products

Genome size vs Total antismash cluster coloured by order

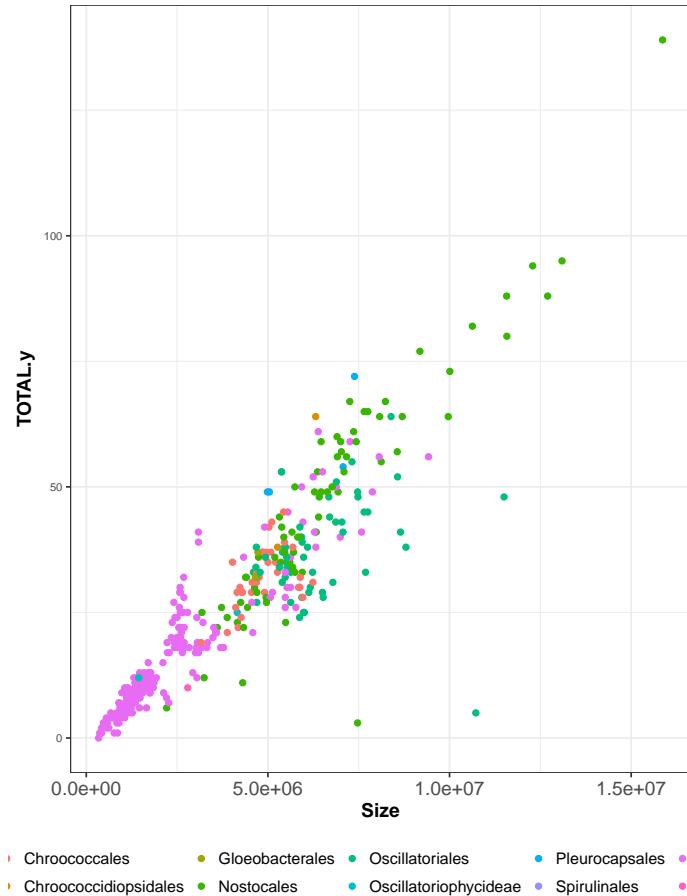


Figure 5.3: Correlation between genome size and antismash Natural products detection colored by Order

Here is a reference to Genome size vs Total antismash cluster: Figure 5.3.

Genome size vs Total antismash cluster detected splitted by order

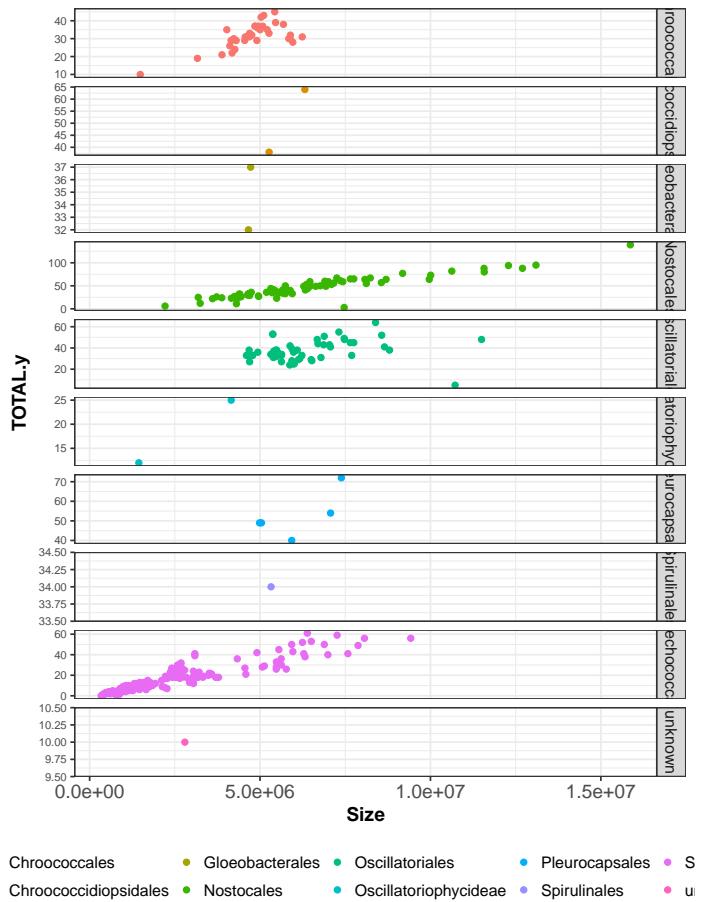


Figure 5.4: Correlation between genome size and antismash Natural products detection grided by Order

Here is a reference to Correlation between genome size and antismash Natural products detection grided by Order plot: Figure 5.4.

5.3.2 Correlation between genome size and Central pathway expansions

Genome size vs Total central pathway expansion coloured by order

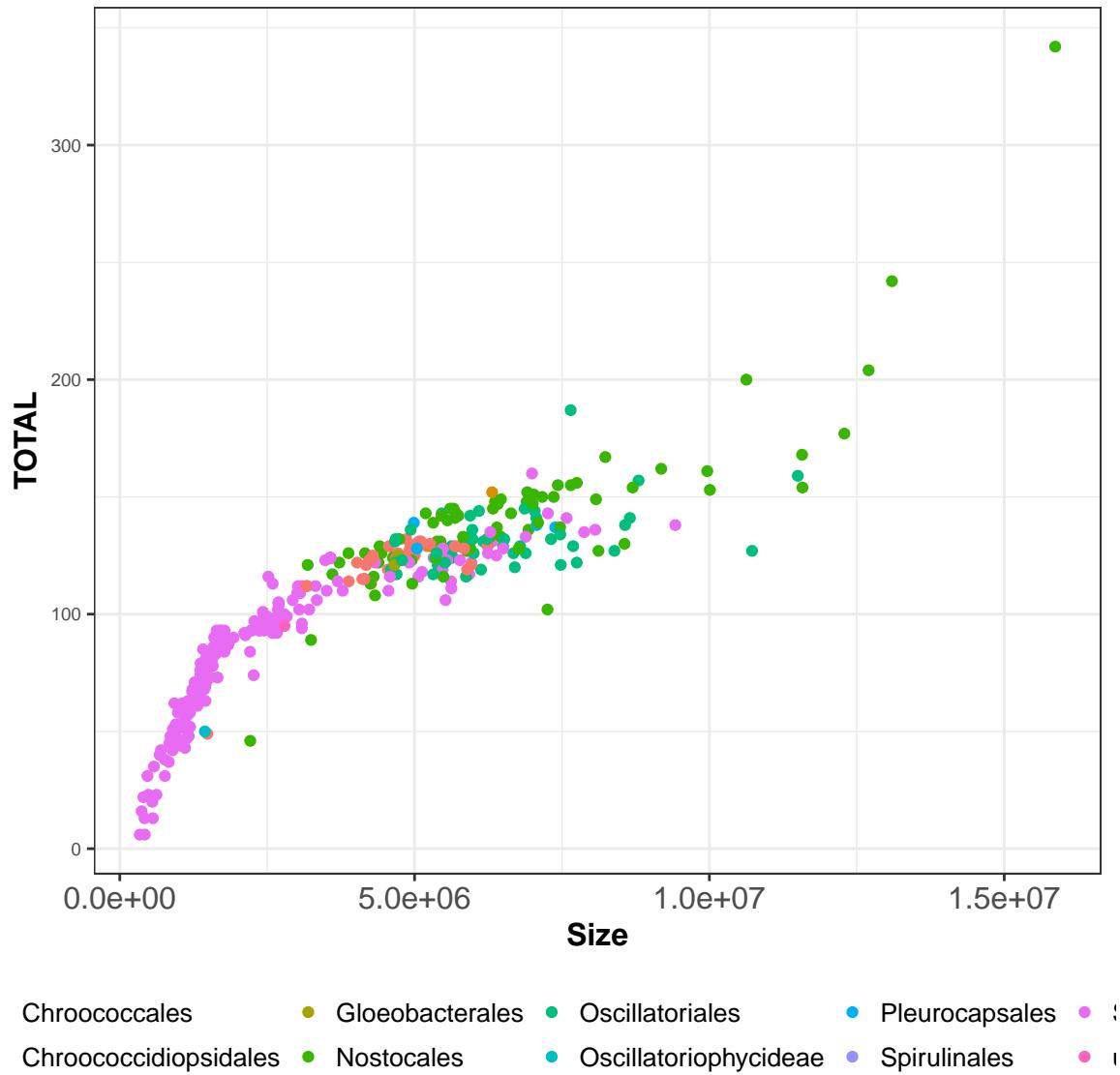


Figure 5.5: Correlation between genome size and central pathway expansions

Here is a reference to the size vs Total central pathway expansion plot: Figure 5.5.

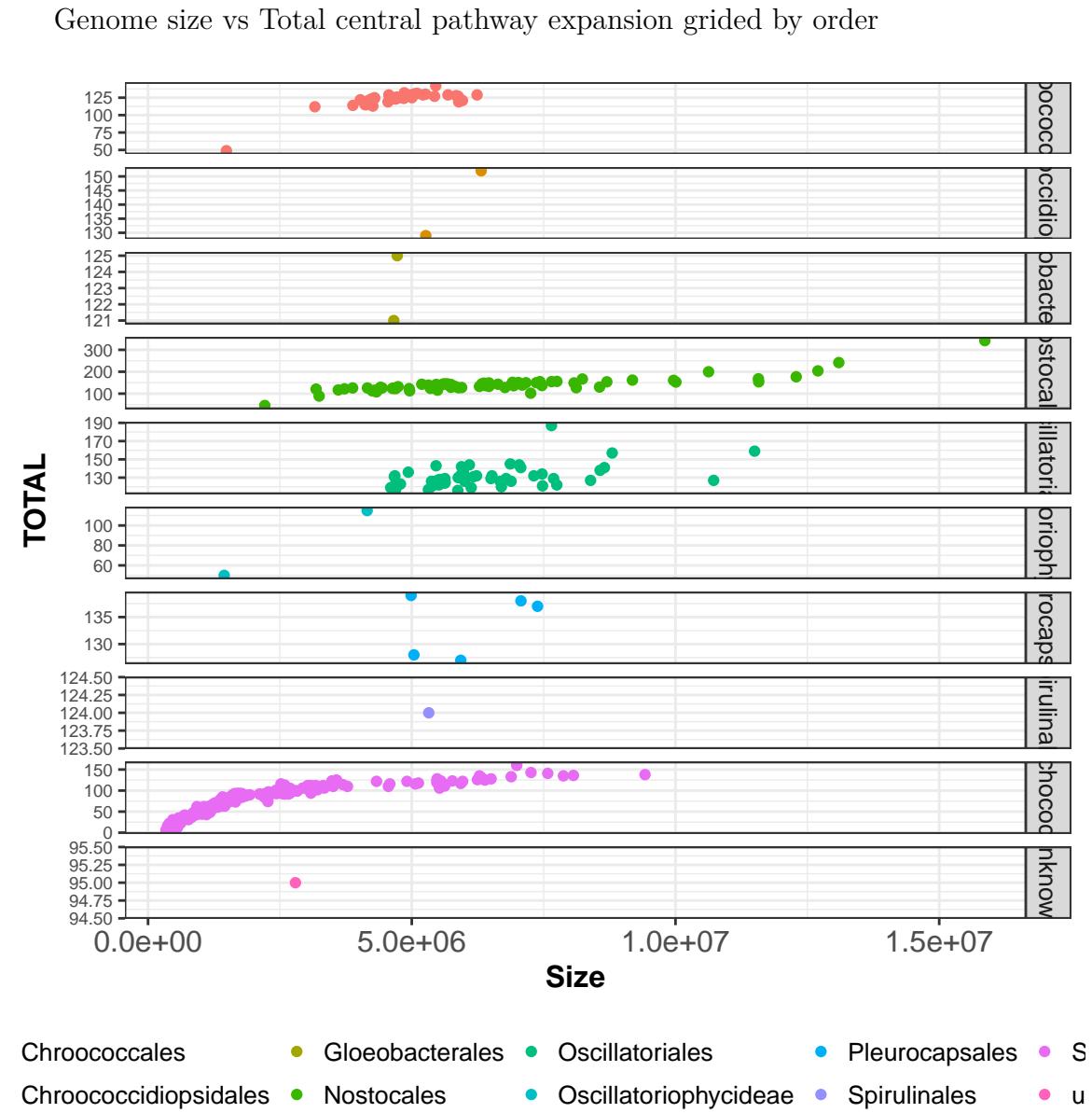


Figure 5.6: Correlation between genome size and central pathway expansions grided by order

Here is a reference to the Genome size vs Total central pathway expansion grided by order plot: Figure 5.6.

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

Also I want to add form given by taxonomical order.

Warning: The shape palette can deal with a maximum of 6 discrete values because more than 6 becomes difficult to discriminate; you have 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

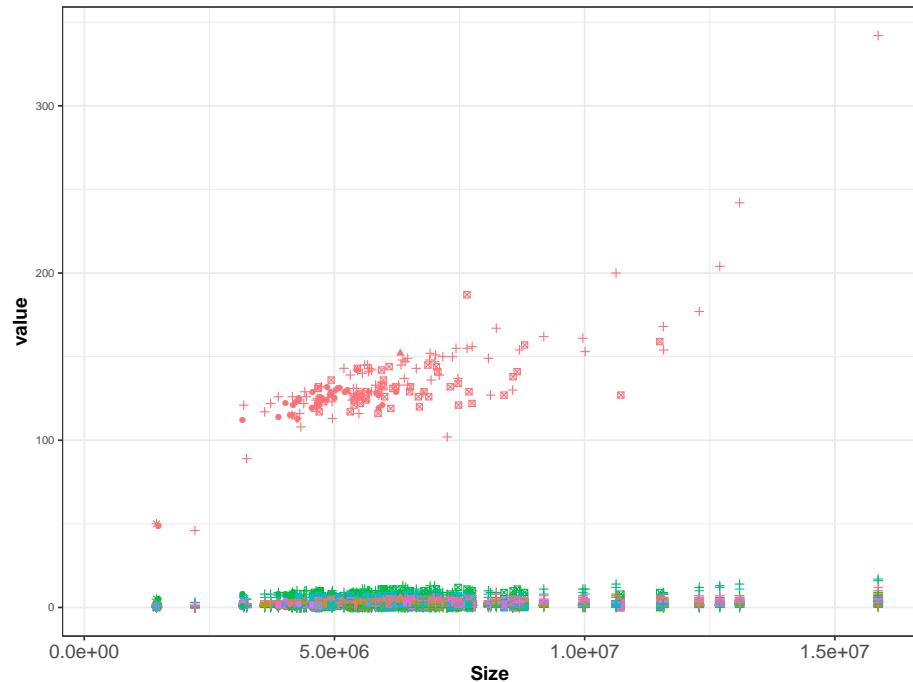


Figure 5.7: Correlation between Genome size vs Total central pathway expansion coloured by metabolic Family

Here is a reference to the Genome size vs Total central pathway expansion coloured by metabolic Family plot: Figure 5.7.

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: ??.

5.4 Natural products

5.4.1 Natural products recruitments from EvoMining heat-plot

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

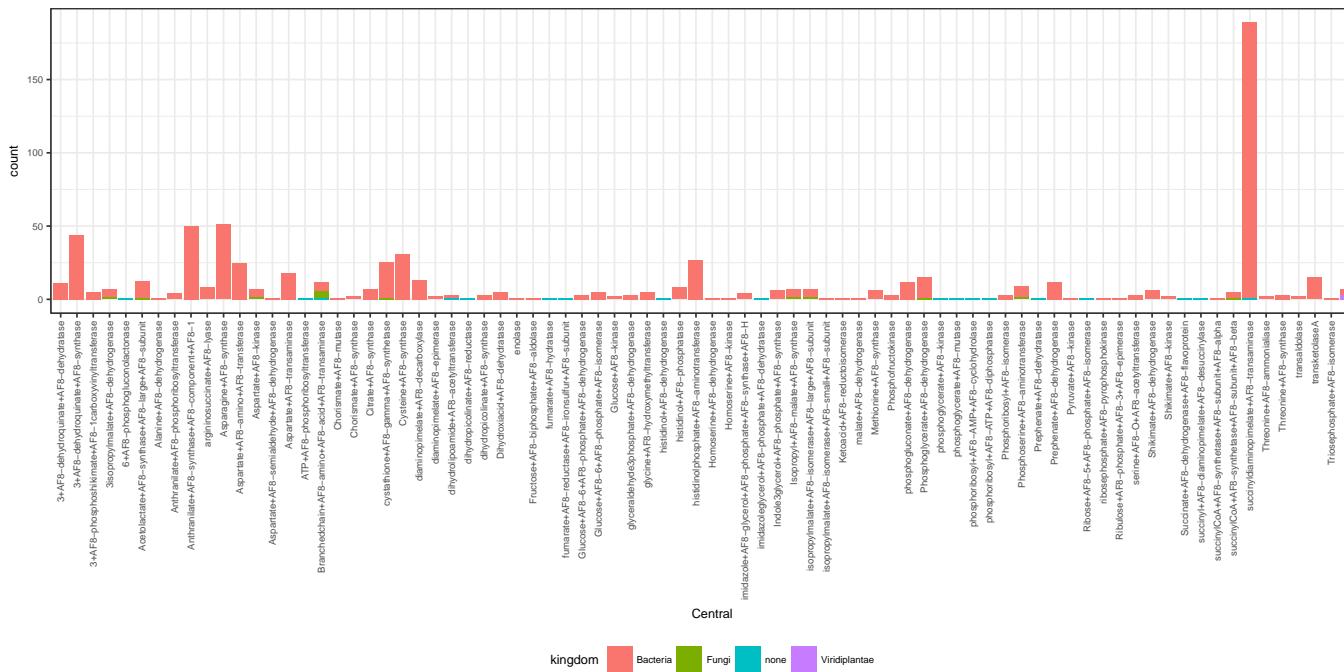


Figure 5.8: Recruitments on central families coloured by kingdom

Here is a reference to Recruitments after central pathways expansions coloured by Kingdom plot: Figure 5.8.

Recruitments after central pathways expansions colourd by taxonomy

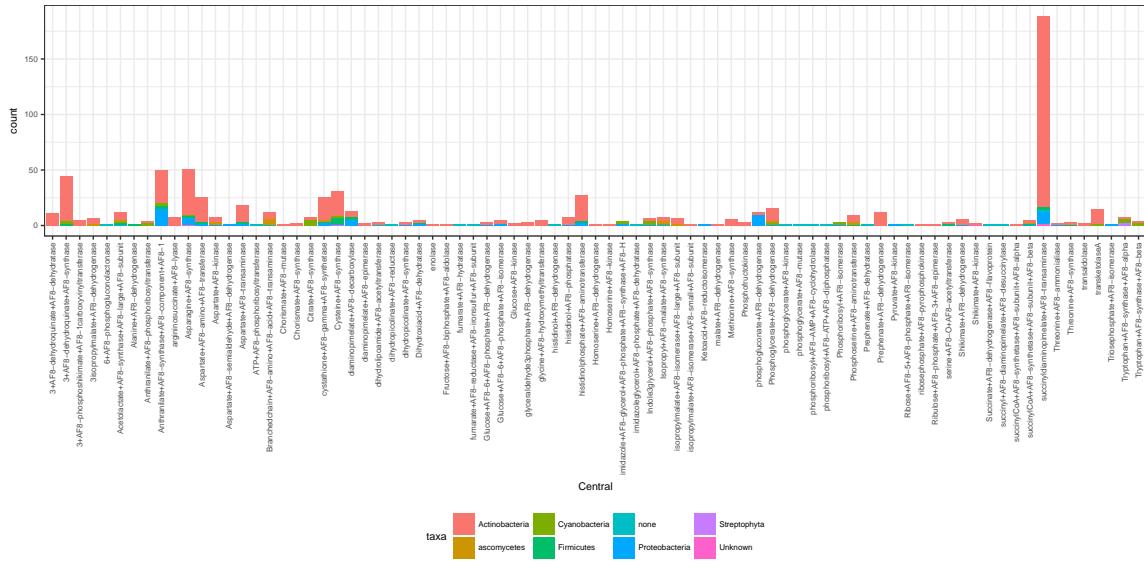


Figure 5.9: Recruitmens on central families coloured by taxonomy

Here is a reference to Recruitments after central pathways expansions coloured by taxa plot: Figure 5.9.

5.5 Cyanobacterias AntiSMASH

Taxonomical diversity on Cyanobacteria Data

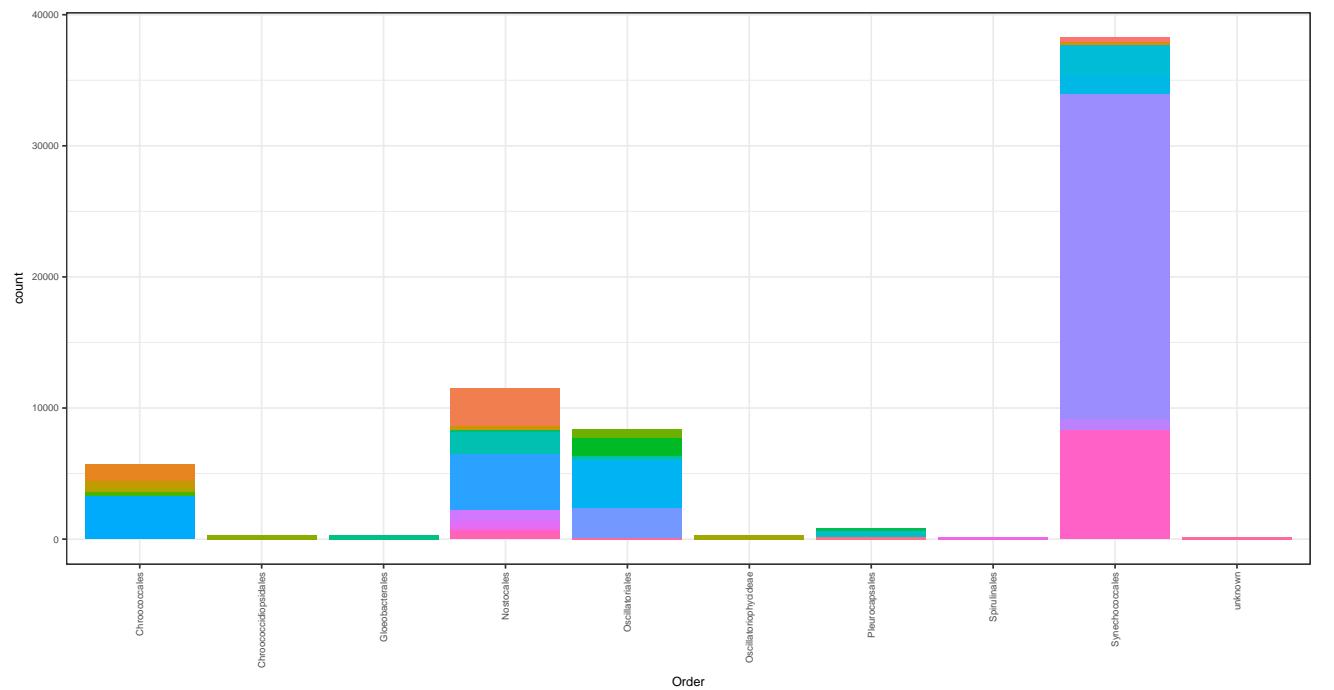


Figure 5.10: Diversity

Here is a reference to Recruitments after central pathways expansions coloured by taxa plot: Figure 5.10.

Smash diversity

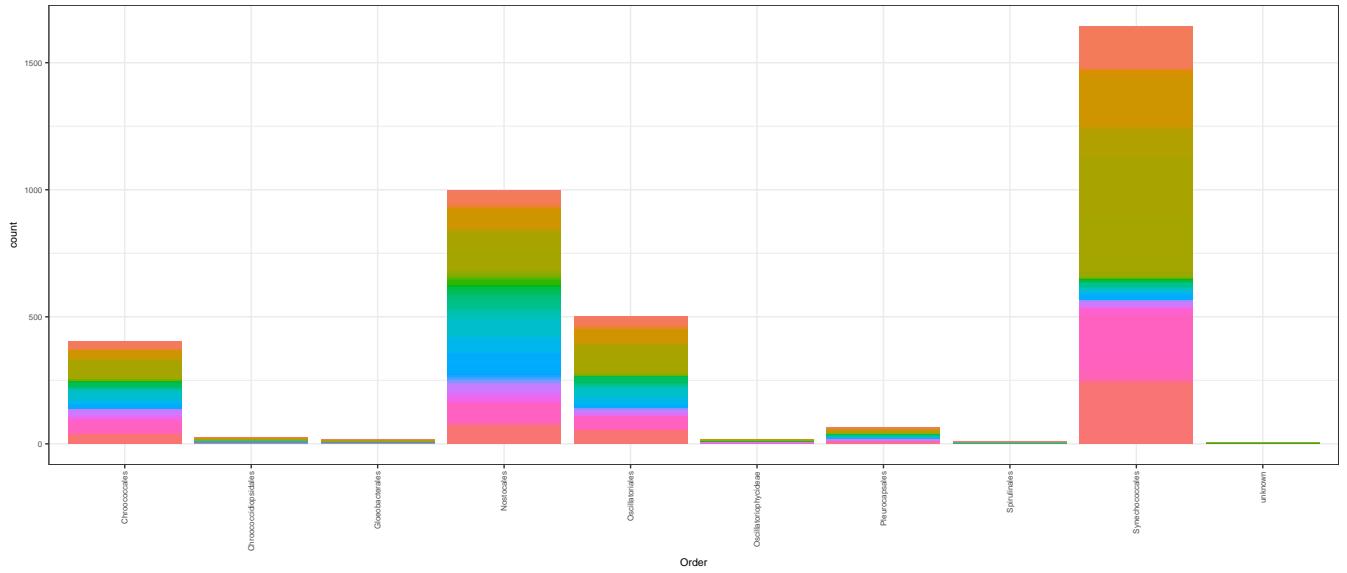


Figure 5.11: Smash

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

5.5.1 AntisMASH vs Central Expansions

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

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

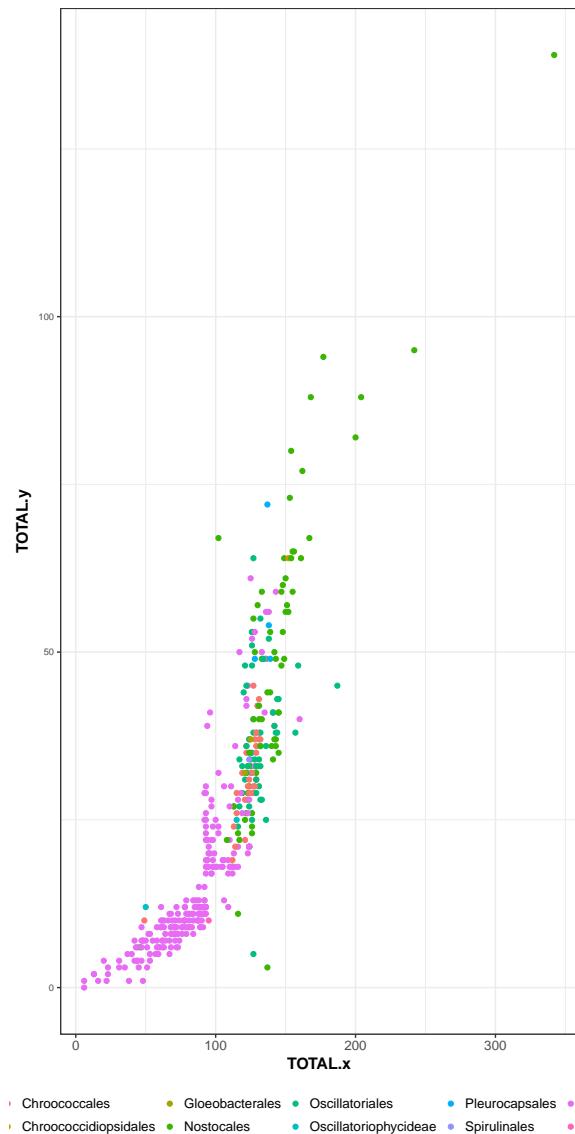


Figure 5.12: Correlation between central pathway expansions and anti-smash Natural products detection

Here is a reference to the expansions vs antismash NP's clusters plot: Figure 5.12.

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

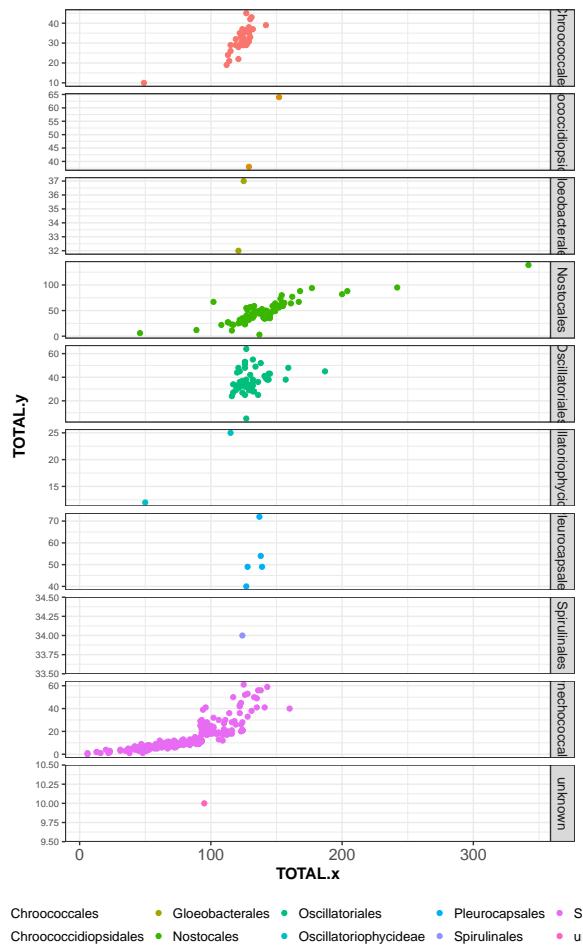


Figure 5.13: Correlation between central pathway expansions and anti-smash Natural products detection

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

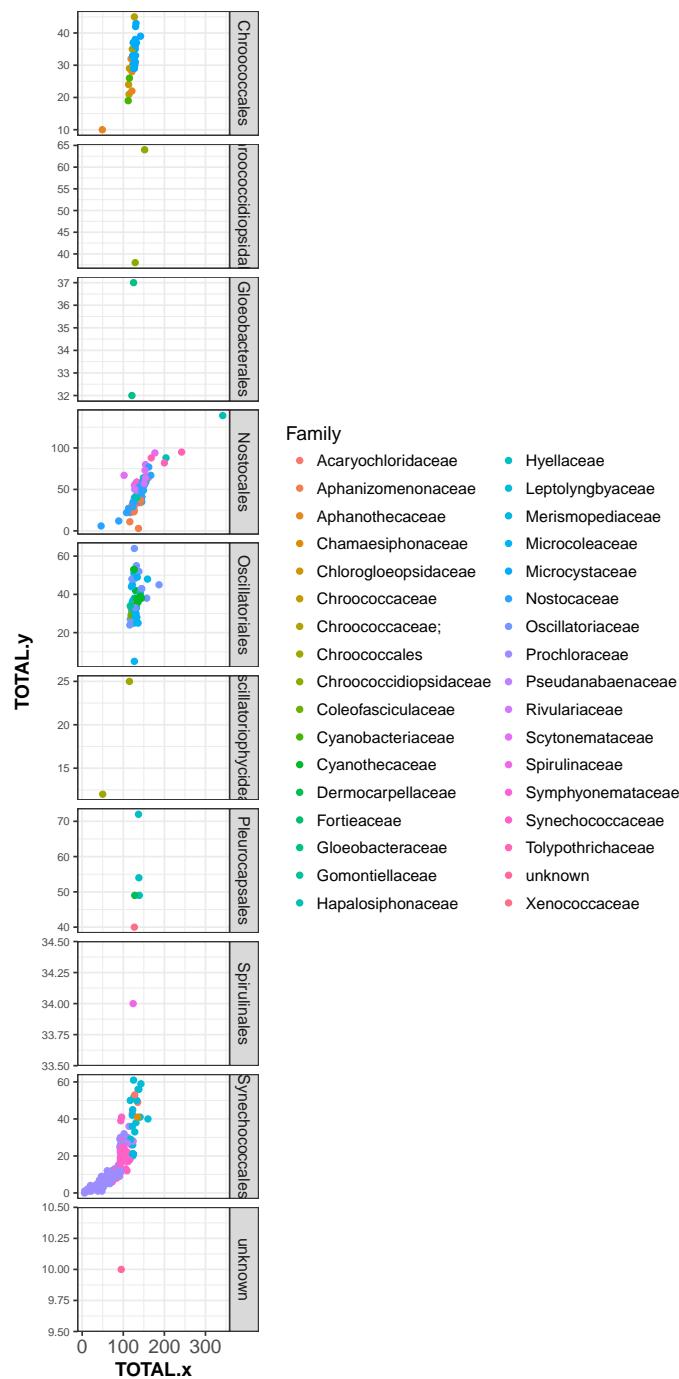


Figure 5.14: Natural products by family

Here is a reference to the Natural products colured by family plot Figure 5.14.

5.6 Selected trees from EvoMining

Phosphoribosyl_isomerase_3 family

Figure from EvoMining

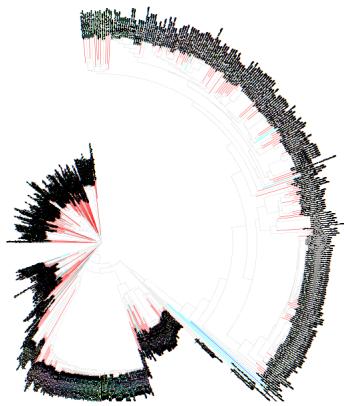


Figure 5.15: Phosphoribosyl isomerase EvoMiningtree

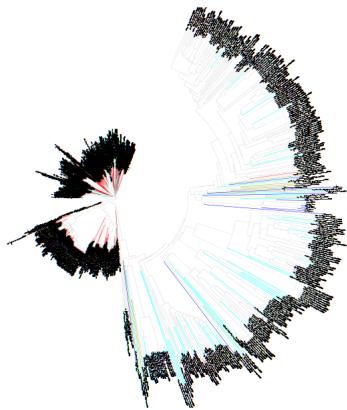


Figure 5.16: Phosphoglycerate dehydrogenase EvoMiningtree

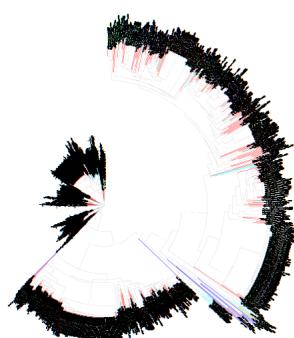


Figure 5.17: Phosphoserine aminotransferase EvoMiningtree

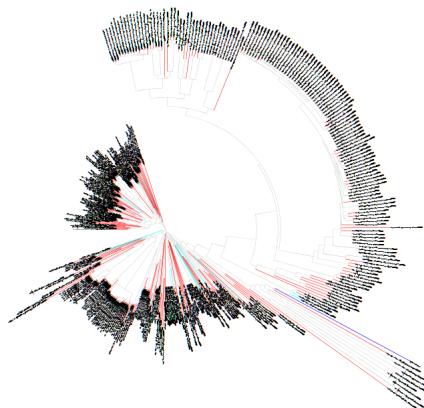


Figure 5.18: Triosephosphate isomerase EvoMiningtree

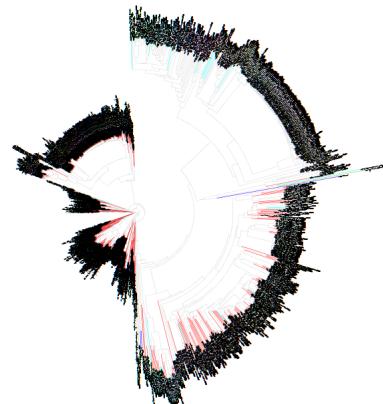


Figure 5.19: glyceraldehyde3phosphate dehydrogenase EvoMiningtree

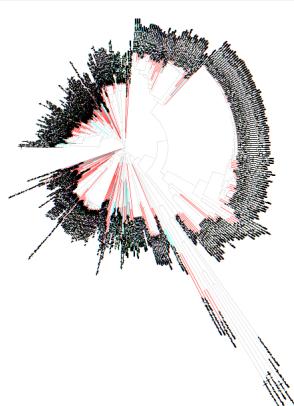


Figure 5.20: phosphoglycerate kinase EvoMiningtree

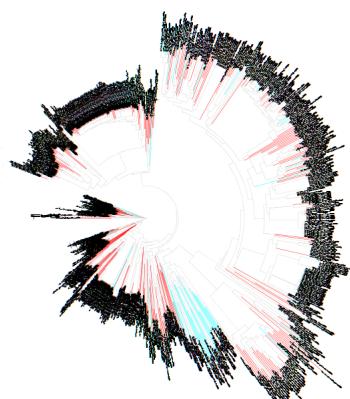


Figure 5.21: phosphoglycerate mutaseEvoMiningtree

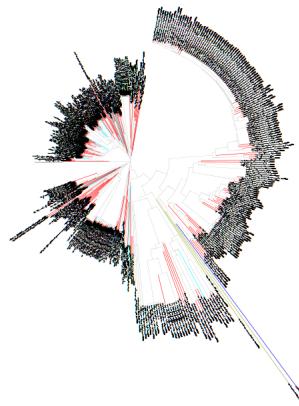


Figure 5.22: enolase EvoMiningtree

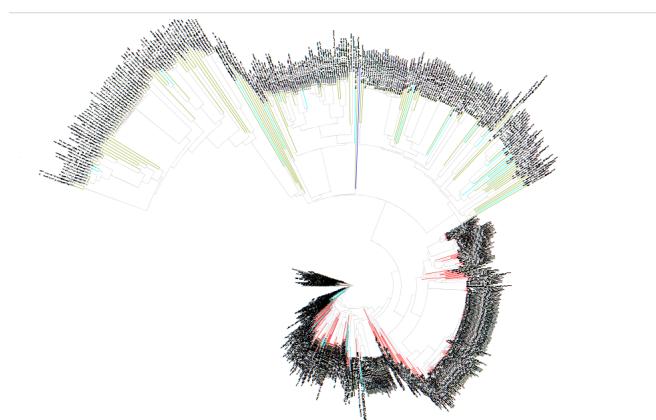


Figure 5.23: Pyruvate kinase EvoMiningtree

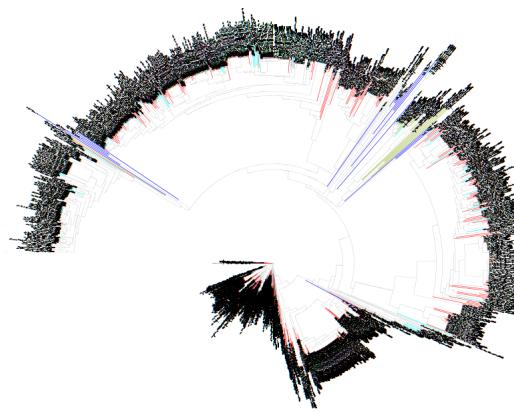


Figure 5.24: Aspartate transaminase EvoMiningtree

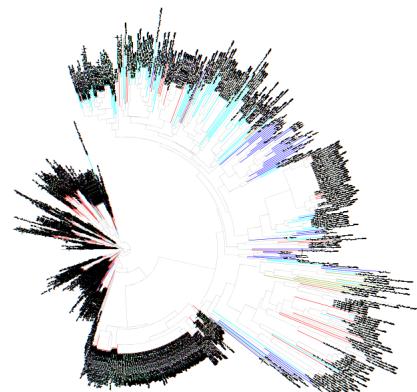


Figure 5.25: Asparagine synthase EvoMiningtree

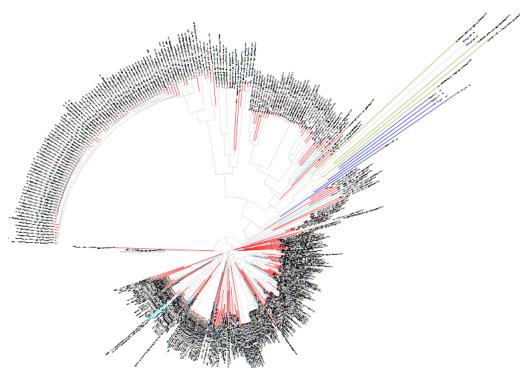


Figure 5.26: Aspartate kinase EvoMiningtree

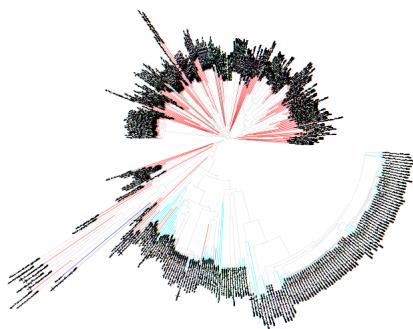


Figure 5.27: Aspartate semialdehyde dehydrogenase EvoMiningtree

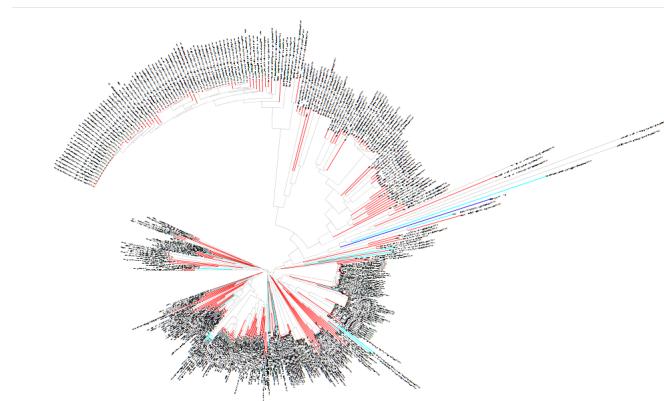


Figure 5.28: Homoserine dehydrogenase EvoMiningtree

Conclusion

Idea de Rosario -ver dell cluster de saxitoxin cuantos pasos se necesitron para llegar ahi.

- A donde se iria el resultado de abrir el GMP
- Otra vez, que Actinos tienen FolE

If we don't want Conclusion to have a chapter number next to it, we can add the `{.unnumbered}` attribute. This has an unintended consequence of the sections being labeled as 3.6 for example though instead of 4.1. The L^AT_EX commands immediately following the Conclusion declaration get things back on track.

More info

And here's some other random info: the first paragraph after a chapter title or section head *shouldn't be* indented, because indents are to tell the reader that you're starting a new paragraph. Since that's obvious after a chapter or section title, proper typesetting doesn't add an indent there.

Appendix A

The First Appendix

This first appendix includes all of the R chunks of code that were hidden throughout the document (using the `include = FALSE` chunk tag) to help with readability and/or setup.

In the main Rmd file:

```
# This chunk ensures that the reedtemplates package is
# installed and loaded. This reedtemplates package includes
# the template files for the thesis and also two functions
# used for labeling and referencing
if(!require(devtools))
  install.packages("devtools", repos = "http://cran.rstudio.com")
if(!require(reedtemplates)){
  library(devtools)
  devtools::install_github("ismayc/reedtemplates")
}
library(reedtemplates)
```

In :

```
# This chunk ensures that the reedtemplates package is
# installed and loaded. This reedtemplates package includes
# the template files for the thesis and also two functions
# used for labeling and referencing
if(!require(devtools))
  install.packages("devtools", repos = "http://cran.rstudio.com")
if(!require(plyr))
  install.packages("plyr", repos = "http://cran.rstudio.com")
if(!require(dplyr))
  install.packages("dplyr", repos = "http://cran.rstudio.com")
```

```
if(!require(ggplot2))
  install.packages("ggplot2", repos = "http://cran.rstudio.com")
if(!require(reedtemplates)){
  library(devtools)
  devtools::install_github("ismayc/reedtemplates")
}
library(reedtemplates)
flights <- read.csv("data/flights.csv")
```

Appendix B

The Second Appendix, for Fun

References

- Angel, E. (2000). *Interactive computer graphics : A top-down approach with openGL*. Boston, MA: Addison Wesley Longman.
- Angel, E. (2001a). *Batch-file computer graphics : A bottom-up approach with quick-Time*. Boston, MA: Wesley Addison Longman.
- Angel, E. (2001b). *Test second book by angel*. Boston, MA: Wesley Addison Longman.
- Molina, S. T., & Borkovec, T. D. (1994). The Penn State worry questionnaire: Psychometric properties and associated characteristics. In G. C. L. Davey & F. Tallis (Eds.), *Worrying: Perspectives on theory, assessment and treatment* (pp. 265–283). New York: Wiley.
- Noble, S. G. (2002). *Turning images into simple line-art* (Undergraduate thesis). Reed College.
- Reed College. (2007, March). LaTeX your document. Retrieved from <http://web.reed.edu/cis/help/LaTeX/index.html>
- Woese, C. R., & Fox, G. E. (1977). Phylogenetic structure of the prokaryotic domain: The primary kingdoms. *Proceedings of the National Academy of Sciences of the United States of America*, 74(11), 5088–5090. Retrieved from <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC432104/>