

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 Selim

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

<b>Background . . . . .</b>	<b>1</b>
0.1 Introduction . . . . .	2
0.1.1 Relacion del pangenoma con la promiscuidad enzimatica . . .	4
0.1.2 Expansion y contextos genomicos como herramienta de anotacion funcional . . . . .	4
0.1.3 Modelos bioinformaticos de promiscuidad . . . . .	5
0.1.4 Promiscuidad in vitro y promiscuidad in vivo . . . . .	6
0.1.5 El papel de la dinamica molecular en la promiscuidad . . . . .	7
0.1.6 Modelo biologico diversidad de Actinobacteria . . . . .	7
0.1.7 Modelo metabolico biosintesis de aminoacidos. . . . .	8
0.2 Antecedentes . . . . .	8
0.2.1 Modelo biologico . . . . .	9
0.2.2 Subsistemas metabolicos . . . . .	9
0.2.3 Contexto y vecindades genomicas . . . . .	9
0.2.4 Metodos bioinformaticos . . . . .	10
0.2.5 Evomining . . . . .	10
0.2.6 Caracterizacion in vivo . . . . .	11
0.2.7 Caracterizacion bioquimica in vitro. . . . .	11
0.2.8 Modelado de dinamica molecular . . . . .	11
0.3 Objetivo General . . . . .	12
0.4 Objetivos particulares . . . . .	12
0.5 Estrategias . . . . .	13
0.5.1 La promiscuidad en familias enzimaticas. . . . .	13
Obtener informacion genomica del phylum Actinobacteria. . .	13
Promiscuidad in vitro dentro de miembros de una familia promiscua de enzimas. . . . .	13
Seleccionar miembros homologos de la familia de enzimas. . .	13
Determinar posibles correlaciones entre los datos producidos..	14
0.5.2 Desarrollar una metodologia para la deteccion in vivo de promiscuidad enzimatica. . . . .	14
Crear cepas geneticamente modificadas con variantes funcionales no nativas de PriA y enzimas asociadas. . . . .	14
0.6 Metodologia . . . . .	15
0.6.1 La promiscuidad en familias enzimaticas. . . . .	15

Actinobacteria genomic . . . . .	15
Annotation . . . . .	15
Genomic DB phylogeny . . . . .	15
0.6.2 Identificar cambios en la vecindad genómica en familias selectas de enzimas de metabolismo central. . . . .	16
Organizar y presentar los datos en una plataforma. . . . .	17
0.6.3 Promiscuidad in vitro . . . . .	18
Datos cinéticos: . . . . .	18
Dinámica molecular . . . . .	18
0.6.4 Promiscuidad in vivo . . . . .	18
0.6.5 Consideraciones . . . . .	19
<b>Chapter 1: EvoMining . . . . .</b>	<b>23</b>
1.1 Introduction . . . . .	23
1.2 Gen families expansions on genomes . . . . .	23
1.2.1 Pangenomes . . . . .	23
1.3 EvoMining . . . . .	23
1.4 Pangenome . . . . .	24
1.5 EvoMining Implementation . . . . .	24
1.6 EvoMining Databases . . . . .	25
Genome DB . . . . .	26
Phylogeny . . . . .	26
Central DB . . . . .	27
1.7 Data Bases . . . . .	27
1.7.1 Central pathways . . . . .	27
1.7.2 Genome Dynamics . . . . .	27
Natural Products DB . . . . .	29
1.7.3 AntisMASH optional DB . . . . .	29
1.7.4 Otras estrategias para los clusters Argon context Idea . . . . .	29
1.8 Argonne . . . . .	29
1.8.1 Inline code . . . . .	32
1.9 Recomendaciones de Luis . . . . .	32
1.10 CORASON: Other genome Mining tools context-based . . . . .	33
1.11 CORe Analysis of Syntenic Orthologs to prioritize Natural Product-Biosynthetic Gene Cluster . . . . .	33
1.12 Loading and exploring data . . . . .	35
<b>Chapter 2: PriA Family . . . . .</b>	<b>39</b>
2.1 Enzymes . . . . .	40
2.2 CORASON . . . . .	40
2.3 Structures . . . . .	41
2.4 Docking . . . . .	42
2.5 Latex no Docking . . . . .	54
2.6 Chemistry 101: Symbols . . . . .	56
2.6.1 Typesetting reactions . . . . .	57

2.6.2	Other examples of reactions . . . . .	57
2.7	Physics . . . . .	57
2.8	Biology . . . . .	57
<b>Chapter 3: Archaea EvoMining Results</b>	. . . . .	<b>59</b>
3.1	Tables . . . . .	60
3.1.1	Expansions BoxPlot by metabolic family . . . . .	61
3.1.2	Expansions BoxPlot by metabolic family by phylum . . . . .	62
3.2	Central pathway expansions . . . . .	73
3.3	Genome Size correlations . . . . .	74
3.3.1	Correlation between genome size and AntiSMASH products . .	74
3.3.2	Correlation between genome size and Central pathway expansions	76
3.4	Natural products . . . . .	80
3.4.1	Natural products recruitments from EvoMining heatplot . .	80
3.5	Archaeas AntiSMASH . . . . .	82
3.5.1	AntisSMASH vs Central Expansions . . . . .	84
3.6	Selected trees from EvoMining . . . . .	87
3.7	. . . . .	88
3.8	Bibliographies . . . . .	88
3.9	Anything else? . . . . .	89
<b>Chapter 4: Actinobacteria EvoMining Results</b>	. . . . .	<b>91</b>
4.1	Tables . . . . .	91
4.1.1	Expansions BoxPlot by metabolic family . . . . .	92
4.2	Central pathway expansions . . . . .	94
4.3	Genome Size correlations . . . . .	98
4.3.1	Correlation between genome size and AntiSMASH products .	98
4.3.2	Correlation between genome size and Central pathway expansions	100
4.4	Natural products . . . . .	104
4.4.1	Natural products recruitments from EvoMining heatplot . .	104
4.5	Actinos AntiSMASH . . . . .	106
4.5.1	AntisSMASH vs Central Expansions . . . . .	108
4.6	Selected trees from EvoMining . . . . .	111
<b>Chapter 5: Cyanobacteria EvoMining Results</b>	. . . . .	<b>113</b>
5.1	Tables . . . . .	113
5.1.1	Expansions BoxPlot by metabolic family . . . . .	114
5.2	Central pathway expansions . . . . .	116
5.3	Genome Size correlations . . . . .	117
5.3.1	Correlation between genome size and AntiSMASH products .	117
5.3.2	Correlation between genome size and Central pathway expansions	119
5.4	Natural products . . . . .	123
5.4.1	Natural products recruitments from EvoMining heatplot . .	123
5.5	Cyanobacterias AntiSMASH . . . . .	125
5.5.1	AntisSMASH vs Central Expansions . . . . .	127

5.6 Selected trees from EvoMining . . . . .	130
<b>Conclusion . . . . .</b>	<b>135</b>
More info . . . . .	135
<b>Appendix A: The First Appendix . . . . .</b>	<b>137</b>
In the main Rmd file: . . . . .	137
In : . . . . .	137
<b>Appendix B: The Second Appendix, Open source code on this document . . . . .</b>	<b>139</b>
B.1 R markdown . . . . .	139
B.2 Docker . . . . .	139
B.3 Git . . . . .	140
B.4 Connect GitHub and DockerHub . . . . .	140
B.5 Additional resources . . . . .	140
<b>References . . . . .</b>	<b>143</b>

# List of Tables

1.1	BBH_Organisms . . . . .	27
1.2	WC_Organisms . . . . .	28
1.3	Maximum Delays by Airline . . . . .	36
2.1	Enzyme PDB . . . . .	41
2.2	Substrates . . . . .	42
2.3	Enzymes docking . . . . .	47
3.1	Families on Archaeabacteria . . . . .	60
4.1	Correlation of Inheritance Factors for Parents and Child . . . . .	91
5.1	Families on Cyanobacteria . . . . .	113



# List of Figures

2.1	Substrates 1 . . . . .	43
2.2	Substrates 2 . . . . .	44
2.3	Substrates 3 . . . . .	45
2.4	Substrates 4 . . . . .	46
2.5	Heatplot docking PriA enzyme Family vs substrates . . . . .	52
2.6	Heat Plot PriA Streptomyces vs other substrates . . . . .	55
2.7	Heat Plot TrpF Streptomyces vs other substrates . . . . .	55
2.8	Combustion of glucose . . . . .	57
3.1	Expansions Boxplot . . . . .	61
3.2	Archaeas Heatplot . . . . .	73
3.3	Correlation between Archaeas genome size and antismash Natural products detection colored by Order . . . . .	74
3.4	Correlation between Archaeas genome size and antismash Natural products detection grided by Order . . . . .	75
3.5	Correlation between Archaeas genome size and central pathway expansions . . . . .	76
3.6	Correlation between Archaeas genome size and central pathway expansions grided by order . . . . .	77
3.7	Correlation between Archaeas Genome size vs Total central pathway expansion coloured by metabolic Family . . . . .	78
3.8	Archaeas Recruitmens on central families coloured by kingdom . . . . .	80
3.9	Archaeas Recruitmens on central families coloured by taxonomy . . . . .	81
3.10	Archaeas Diversity . . . . .	82
3.11	Archaeas Smash Taxonomical Diversity . . . . .	83
3.12	Correlation between Archaeas central pathway expansions and anti-smash Natural products detection . . . . .	84
3.13	Correlation between Archaeas central pathway expnasions and anti-smash Natural products detection . . . . .	85
3.14	Archaeas Natural products by family . . . . .	86
3.15	Phosphoribosyl isomerase A EvoMiningtree . . . . .	87
3.16	Phosphoribosyl isomerase other EvoMiningtree . . . . .	87
3.17	Phosphoribosyl anthranilate isomerase EvoMiningtree . . . . .	87
4.1	Expansions Boxplot . . . . .	93

4.2	Actinobacterial Heatplot . . . . .	95
4.3	Streptomyces Genomes expansions on PGA Aminoacids HeatPlot . . . . .	97
4.4	Correlation between Actinos genome size and antismash Natural products detection colored by Order . . . . .	98
4.5	Correlation between Actinos genome size and antismash Natural products detection grided by Order . . . . .	99
4.6	Correlation between Actinos genome size and central pathway expansions . . . . .	100
4.7	Correlation between Actinos genome size and central pathway expansions grided by order . . . . .	101
4.8	Correlation between Actinos Genome size vs Total central pathway expansion coloured by metabolic Family . . . . .	102
4.9	Actinos Recruitmens on central families coloured by kingdom . . . . .	104
4.10	Actinos Recruitmens on central families coloured by taxonomy . . . . .	105
4.11	Actinos Diversity . . . . .	106
4.12	Actinos Smash Taxonomical Diversity . . . . .	107
4.13	Correlation between Actinos central pathway expansions and antismash Natural products detection . . . . .	108
4.14	Correlation between Actinos central pathway expnasions and antismash Natural products detection . . . . .	109
4.15	Actinos Natural products by family . . . . .	110
4.16	Enolase EvoMiningtree . . . . .	111
4.17	Phosphoribosyl isomerase EvoMiningtree . . . . .	111
4.18	Phosphoribosyl isomerase A EvoMiningtree . . . . .	111
4.19	phosphoshikimate carboxyvinyltransferase EvoMiningtree . . . . .	112
5.1	Expansions Boxplot . . . . .	115
5.2	Cyanobacterial Heatplot . . . . .	116
5.3	Correlation between genome size and antismash Natural products detection colored by Order . . . . .	117
5.4	Correlation between genome size and antismash Natural products detection grided by Order . . . . .	118
5.5	Correlation between genome size and central pathway expansions . . . . .	119
5.6	Correlation between genome size and central pathway expansions grided by order . . . . .	120
5.7	Correlation between Genome size vs Total central pathway expansion coloured by metabolic Family . . . . .	121
5.8	Recruitmens on central families coloured by kingdom . . . . .	123
5.9	Recruitmens on central families coloured by taxonomy . . . . .	124
5.10	Diversity . . . . .	125
5.11	Smash . . . . .	126
5.12	Correlation between central pathway axpnasions and antismash Natural products detection . . . . .	127
5.13	Correlation between central pathway axpnasions and antismash Natural products detection . . . . .	128

5.14 Natural products by family . . . . .	129
5.15 Phosphoribosyl isomerase EvoMiningtree . . . . .	130
5.16 Phosphoglycerate dehydrogenase EvoMiningtree . . . . .	130
5.17 Phosphoserine aminotransferase EvoMiningtree . . . . .	130
5.18 Triosephosphate isomerase EvoMiningtree . . . . .	131
5.19 glyceraldehyde3phosphate dehydrogenase EvoMiningtree . . . . .	131
5.20 phosphoglycerate kinase EvoMiningtree . . . . .	131
5.21 phosphoglycerate mutaseEvoMiningtree . . . . .	132
5.22 enolase EvoMiningtree . . . . .	132
5.23 Pyruvate kinase EvoMiningtree . . . . .	132
5.24 Aspartate transaminase EvoMiningtree . . . . .	133
5.25 Asparagine synthase EvoMiningtree . . . . .	133
5.26 Aspartate kinase EvoMiningtree . . . . .	133
5.27 Aspartate semialdehyde dehydrogenase EvoMiningtree . . . . .	134
5.28 Homoserine dehydrogenase EvoMiningtree . . . . .	134



# **Abstract**

The preface pretty much says it all.

Second paragraph of abstract starts here.



# **Dedication**

You can have a dedication here if you wish.



# Background

Enzymes catalyze chemical reactions transforming substrates into products. During 20th century enzymes were perceived as highly specific catalysts, nevertheless this perception changed with the discovery that they can . This ability to catalyze several chemical functions is known as enzyme promiscuity. *Escherichia coli* contains at least 404 promiscuous enzymes. La relevancia de la promiscuidad radica tanto en su papel como mecanismo de evolución de la función enzimática, así como en la necesidad de su detección para la corrección de modelos de flujo metabólico y la determinación de efectos secundarios en drogas farmacológicas. A pesar de su frecuencia e importancia aún se está en el proceso de entender las causas y las características observables de la promiscuidad enzimática.

Para estudiar la promiscuidad esta propuesta discierne entre dos problemas. El primero es el de ubicar cuáles familias tienen enzimas promiscuas, al que llamaremos problema de las familias. El segundo es el de los miembros: una vez identificada una familia promiscua, cómo distinguir entre sus miembros enzimas con distinto nivel de promiscuidad. Se ha intentado identificar enzimas promiscuas, a nivel de secuencia, sin pasar por experimentación mediante aprendizaje maquinaria. Estos enfoques son incapaces de identificar una familia promiscua si no se conoce previamente al menos un miembro promiscuo de ella. Por otra parte, en el problema de los miembros se presentan dificultades cuando la identidad de secuencia es alta, e.g. en la familia PriA-HisA se sabe que la enzima HisA de *E. coli* no es promiscua, pero PriA de *Streptomyces coelicolor* sí lo es.

Para mejorar nuestro entendimiento del fenómeno, además de la comparación de secuencias es necesario integrar otros elementos de análisis. Se debe notar que es prácticamente imposible decir que una enzima no es promiscua ya que para ello se deberían haber descartado todos los posibles sustratos. Sin embargo, para el estudio de cambios en promiscuidad se han detectado como elementos relevantes a los cambios en vecindad genómica, los cambios en flexibilidad durante la dinámica molecular, la pérdida de genes centrales, y finalmente a las expansiones genéticas dentro de un grupo taxonómico. Estos elementos tienen en común que reflejan un cambio en alguna propiedad genómica o biofísica, de lo que se deriva que el buscar cambios en la promiscuidad de una enzima resulta más factible que la búsqueda intrínseca de promiscuidad, a lo que aspiran los métodos basados en comparaciones de secuencias.

Evomining es una plataforma bioinformatica pensada para la busqueda de expansiones de familias genicas de metabolismo central. Desarrollarla en combinacion con algoritmos de busqueda de cambios en la vecindad genomica la haran una plataforma ideal para abordar el problema de las familias, proporcionando una solucion a la dificultad de no tener conocimiento previo de un miembro promiscuo en la familia investigada. Respecto al problema de los miembros, se propone explorar variaciones en vecindad genomica, flujo genico y dinamica molecular, como candidatos a reflejar la variacion en promiscuidad. Finalmente, he detectado que a pesar de que pruebas *in vivo* son mas sensibles a niveles bajos de promiscuidad que mediciones *in vitro*, esta ultima suele ser la unica estudiada. *In vivo*, la metabolomica aplicada en genes biosinteticos de productos naturales ha ayudado en la identificacion de sustratos, por lo que esta tecnica podria ayudar a revelar el nuevo sustrato de una enzima en la que se sospecha ganancia de promiscuidad. En resumen, el objetivo de mi trabajo sera abordar los dos problemas de promiscuidad considerando la diferenciacion *in vitro* e *in vivo* tomando como modelo biologico el phylum Actinobacteria, un grupo de bacterias reconocido por su diversidad metabolica donde se ha probado la existencia de promiscuidad enzimatica.

## 0.1 Introduction

Para estudiar la promiscuidad es necesario contar con una definicion, algunos autores emplean el termino promiscuidad para describir actividades enzimaticas distintas a la funcion principal [4] otros lo ven como una actividad secundaria fortuita [5] que pudo aparecer de forma accidental o inducida artificialmente [6]. Otros mas, cuando una enzima puede operar sobre un amplio rango de sustratos, prefieren llamarla multiespecifica [4]. A la accion de realizar distintas funciones cataliticas, ya sea al catalizar varias reacciones quimicas o bien una misma reaccion en sustratos diferentes se le conoce como promiscuidad enzimatica [7]. Existen varios tipos de promiscuidad enzimatica.

Por sustrato cuando la reaccion es la misma pero se lleva a cabo en distintos sustratos ejemplo la familia PriA [8] y la familia de betalactamasas [9]

Catalitica cuando la enzima utiliza diferentes mecanismos de reaccion y/o residuos cataliticos, e.g. la quimotripsina puede catalizar reacciones de amidasa y fosfotriesterasa en un mismo sitio activo. [7] Por condiciones del entorno, cuando la enzima cambia su conformacion dependiendo de las condiciones quimicas y fisicas presentes como pH, temperatura, solventes organicos y salinidad e.g. algunas lipasas pueden actuar como sintetizadoras de esteres en lugar de hidrolasas en presencia de solventes organicos [6].

Este trabajo se enfocara a la promiscuidad por sustrato, entendiendo asi que la enzima es capaz de catalizar la misma reaccion quimica en al menos dos sustratos. La promiscuidad por sustrato es importante en terminos evolutivos, por ejemplo la enzyme commission number (EC) separa las enzimas en clases, a cada enzima se

asignan 4 digitos, los tres primeros corresponden a la reaccion y el ultimo al sustrato; el mayor numero de sustratos (4306 clases) que de reacciones quimicas (234 en el tercer nivel) sugiere que la mayor variacion evolutiva se da a nivel de sustrato y no de reaccion [11]. Otra evidencia de la importancia de la multiespecificidad por sustrato esta en el descubrimiento de las superfamilias, enzimas mecanistica y estructuralmente relacionadas que divergen en su afinidad por sustrato [12]

Si bien existen familias de enzimas con alta especificidad por sustrato, otras familias como el citocromo P450 [14] y las beta lactamasas [16] son promiscuas. Es posible que la vision previa de alta especificidad se deba a que las primeras rutas metabolicas estudiadas pertenecen al metabolismo central, donde la especificidad puede haber sido favorecida por presiones de seleccion [17]. Esta vision ha cambiado debido al conocimiento de mas enzimas con multifuncionalidad [18], sin afectar la eficiencia catalitica por la funcion primaria [19]. En 1976 el interes por la promiscuidad comenzó por su influencia en la evolucion de la funcion enzimática[20], las aproximaciones variaron desde la aparicion de la sintesis funcional [22], cuando la disponibilidad de genomas permitio la combinacion de analisis filogeneticos con tecnicas de biologia molecular, bioquimica y biofisica (Fig 1). En 2003 la biofisica de las proteinas entra en escena al postularse que la diversidad conformacional durante la dinamica molecular debe incidir en la aceptacion de distintos sustratos. Recientemente se ha investigado su papel en efectos secundarios en drogas farmacologicas [23]. Entre 2005 y 2010 se avanza del estudio de una sola familia enzimática hacia el interes por propiedades globales, por ejemplo dado un genoma se investiga la distribucion de familias promiscuas en subsistemas metabolicos. En estos años, surge el desarrollo de indices que reflejen las caracteristicas bioquimicas de enzimas promiscuas. En 2010, comienzan los intentos por desarrollar un metodo computacional de predicción de promiscuidad. Desde 2012 a la fecha, a la par que las aproximaciones bioinformaticas se multiplican, se desarrollan investigaciones de aspectos biofisicos, bioquimicos y evolutivos de enzimas promiscuas reafirmando que todos estos aspectos estan relacionados al fenomeno. En las siguientes secciones se describiran trabajos importantes sobre la relacion que guarda la promiscuidad con expansiones genomicas y flexibilidad molecular. Ademas se hablara sobre analisis bioquimicos y metabolicos para la descripcion del fenomeno.

TeX or L<sup>A</sup>T<sub>E</sub>X

###Funcion biologica de la promiscuidad enzimática

¿Por que existe la promiscuidad enzimática? Se tiene evidencia de dos papeles biologicos: el primero proporcionar robustez a la red metabolica de un organismo mediante redundancia de reacciones de otras enzimas; el segundo permitir plasticidad evolutiva, es decir materia prima para la adaptacion a variaciones ambientales [19] mediante la adquisicion de nuevas funciones quimicas. Respecto a la robustez, se probó que sobreexpresar enzimas promiscuas puede rescatar perdidas genicas [30]. De 104 knockout sencillos de genes esenciales para *E. coli* K-12, 20% de las auxotrofias pudieron ser suprimidas por la sobreexpresión de plasmidos que contenian enzimas promiscuas. Otro ejemplo que aporta a la robustez es PriA, enzima de la ruta de histidina que realiza en la ruta del triptofano la reaccion E.C. 5.3.1.24 [8]. En cuanto a la plasticidad se propone que para que la promiscuidad pueda dar origen a la aparicion de nuevas

funciones la actividad promiscua debe proveer una ventaja fisiologica inmediata para poder ser seleccionada positivamente, ademas una vez que una funcion promiscua se vuelva relevante se debe poder mejorar mediante pocas mutaciones derivando en el intercambio entre la actividad promiscua y la principal[4].

Aun cuando el producto de la promiscuidad genera metabolitos que no se integran al metabolismo central de la celula, su efecto es positivo ya que estos metabolitos podrian colaborar a la adaptacion al entorno participando por ejemplo en una relacion de simbiosis o de competencia con otros organismos. Este tipo de metabolitos, por lo general, no son dañinos [31] y pueden servir como bloques de construccion para vias metabolicas nuevas [34]. La respuesta inmediata de adaptacion de un organismo podria ser una consecuencia de su grado de promiscuidad.

### **0.1.1 Relacion del pangenoma con la promiscuidad enzimatica**

El core genome de un grupo taxonomico es el conjunto de secuencias codificantes presentes en todos los organismos del grupo. En el Dominio Bacteria el core esta estimado entre 200 y 300 secuencias [37]. Dada su conservacion el core genome puede utilizarse para trazar mejores relaciones filogeneticas que las obtenidas con el uso exclusivo de marcadores como la subunidad 16s del RNA ribosomal o el gen rpoB. El pangenoma es el conjunto complemento del core genome, es decir todas aquellas secuencias que estan ausentes de uno o mas organismos del grupo y por lo tanto no son necesarias para todos, sino solo posiblemente para el organismo que las posee. Como en el pangenoma la presion de seleccion esta relajada respecto al core-genome [17] es el conjunto donde la plasticidad genomica tiene facilidades para desarrollarse.

Esta idea puede restringirse a subsistemas metabolicos para identificar genes cuyas enzimas estan en proceso de cambio de funcion quimica, por ejemplo, en este trabajo se encontro que el gen *trpF* esta presente en solo 49 de 290 genomas analizados del genero Streptomyces por lo que se encuentra en el pangenoma de triptofano de este genero taxonomico, posiblemente adquiriendo una nueva funcion [34]. Para evitar problemas tecnicos del calculo del pangenoma existen otros modelos de medicion de variabilidad del genomica entre especies bacterianas [38].

### **0.1.2 Expansion y contextos genomicos como herramienta de anotacion funcional**

La diversidad enzimática existente es el resultado de un proceso de expansion, mutacion y seleccion que se ha desarrollado durante el transcurso de la historia evolutiva [4]. Existe evidencia de que cierto grado de promiscuidad o divergencia funcional precede a la duplicacion genica [40]. Por este motivo detectar expansiones ya sea duplicaciones o transferencias horizontales [41], puede ser un buen punto de partida para determinar

divergencia funcional y promiscuidad. No todas las expansiones denotan cambio de funcion enzimatica, algunas pueden ser meros accidentes, sin embargo dado que la funcion de una enzima suele estar relacionada con sus vecinos [42], una expansion en una vecindad genomica diferente de la tradicional sera un referente de adquisicion de una nueva funcion y entonces un indicador de existencia previa de promiscuidad.

La funcion de una enzima es un concepto jerarquico, dependiente de la filogenia de un organismo [46]. Para sistematizar el estudio de contextos y vecindades genomicas se desarrollo Search Tool for the Retrieval of Interacting Genes/Proteins STRING [47], que cuenta con una anotacion de ortologia jerarquica y consistente, realizada en 2000 organismos en cuyo marco interacciones de proteinas con implicaciones funcionales son predichas tanto de novo por informacion genomica de co-ocurrencia como por mineria de datos en articulos publicados. STRING es una base de datos, y como tal no permite agregar nuevos genomas para su analisis. Sus 2000 organismos incluyen especies tanto bacterianas como eucariotas. Al existir tanta diversidad, los genomas disponibles para un genero o clase especificos son escasos, p. g. de los mas de 300 genomas disponibles de Streptomyces solo 24 estan incluidos.

Para resolver la baja cobertura de STRING hacia ciertos grupos taxonomicos se pueden desarrollar scripts de vecindad genomica utilizando RAST (Rapid Annotation using Subsystem Technology); un servicio interactivo de anotacion automatica de genomas de bacterias y arqueas [48] donde la funcion de cada gen se asigna de acuerdo a conocimiento previo de subsecuencias de organismos cercanos filogeneticamente, cuando es posible se incluye en un subsistema metabolico. Estamos en una era de explosion de datos genomicos, proximamente se espera contar con millones de genomas bacterianos incluso provenientes de bacterias no cultivables, por ello los algoritmos deben ser constantemente optimizados a los nuevos volumenes de datos [50]. Ante esta expectativa seria muy util desarrollar algoritmos de analisis genomico que sean de codigo libre o al menos interactivos para que cada laboratorio pueda personalizarlos para sus propios genomas.

Finalmente, no solo la vecindad genomica inmediata puede ser utilizada como distinutivo en la busqueda de promiscuidad, diferencias en el contexto genomico en genes relacionados con una enzima promiscua, sin importar su ubicacion dentro del genoma tambien pueden ser relevantes para la perdida o ganancia de funcion quimica [51], (Juarez Vazquez et al 2015).

### 0.1.3 Modelos bioinformaticos de promiscuidad

Con el fin de reducir la inversion en el proceso de experimentacion, se han implementado en los ultimos años algoritmos computacionales para predecir promiscuidad enzimatica [52]. Estos procedimientos cuentan con un conjunto de aprendizaje, unos descriptores del conjunto, una fase de ajuste de parametros y finalmente una prediccion. En 2010, Carbonell propone un algoritmo de soporte vectorial basado en subsecuencias de distinto tamaño que llama huellas moleculares. En este trabajo aplicado sobre 500,000

proteinas reportadas en la enciclopedia de Kyoto de genes y genomas (KEGG) se reporta 85% de exito en detección de enzimas promiscuas anotadas en KEGG. En 2012, Cheng compara los métodos de random forest y soporte vectorial en 6799 proteínas provenientes de la base de datos Universal Protein Resource (UniProt). Las enzimas son descritas con subsecuencias de aminoácidos incorporando además características biofísicas como polaridad. Se utiliza como grupo de control a familias de enzimas donde nunca se ha reportado una enzima promiscua.

Un aspecto no considerado en estos métodos es que hay familias de enzimas con alta identidad de secuencia entre sus miembros, con cambios bruscos en promiscuidad, debidos por ejemplo a la dinámica genómica [51], lo que dificulta que considerar solo la secuencia lleve a buenos predictores de promiscuidad. Cuando se obtiene una predicción positiva utilizando los modelos existentes, lo que significa es que dada esa secuencia, en su familia se conoce previamente un elemento promiscuo y que además sus subsecuencias de cierto tamaño son suficientemente similares. Estos enfoques no pueden predecir de novo, en familias donde la promiscuidad no ha sido previamente detectada experimentalmente, pues no consideran aspectos evolutivos ni mecanísticos de las enzimas.

Otra limitante a los enfoques descritos es que mezclan en su conjunto de entrenamiento fenómenos distintos de promiscuidad. Cheng p. g. incluye enzimas moonlight que si bien poseen funciones adicionales a la catalización, son distantes a las enzimas promiscuas [5]. Además en ambos casos mezclan en el mismo conjunto enzimas bacterianas y eucariotas, con lo que si existía una huella basada en secuencia entonces esta puede diluirse por la gran distancia taxonómica entre estos grupos (Tabla 1).

#### 0.1.4 Promiscuidad *in vitro* y promiscuidad *in vivo*

La ganancia de promiscuidad no solo puede entenderse como la capacidad de convertir más sustratos [52], sino también como la mejora de la capacidad catalítica respecto a ellos. El I-index [15], está definido como un rango de valores entre 0 y 1 que tiende a 1 entre más parecida sea la actividad de la enzima sobre distintos sustratos, la capacidad catalítica es medida en términos del cociente de Michaelis - Menten  $\frac{K_{cat}}{K_m}$ . El índice ha sido utilizado para predecir la afinidad por sustrato del citocromo P450 [25]. Una limitante del índice *I* es que se deben conocer los sustratos a los que la enzima es afin; sin embargo se puede sospechar que una enzima ha ganado promiscuidad aun sin conocer sus potenciales sustratos. Otro punto a señalar es que las variables *K<sub>cat</sub>*, *K<sub>m</sub>* son mediciones realizadas *in vitro* y no se consideran todos los sustratos presentes *in vivo*. Para solventar esta dificultad e investigar variaciones de sustratos nativos se pueden buscar productos similares a los ya conocidos por medio de análisis metabólicos [57] como los empleados en la detección de rutas no conservadas en la biosíntesis de productos naturales [50]. En particular para este fin se ha utilizado espectrometría de masas MS/MS, [57] combinada con molecular networking para identificar productos similares [59]

### 0.1.5 El papel de la dinamica molecular en la promiscuidad

La estructura tridimensional de una proteina es obtenida mediante previa purificacion y cristalizacion. Aunque mucho se ha hablado de la relacion estructura funcion, al cristalizar se obtienen estados conformacionales homogeneos, que bien pueden no ser la unica conformacion que adopta la proteina en solucion. [61]. En particular en el problema de promiscuidad, se ha observado que la variacion funcional no queda obviamente reflejada en la variacion estructural, lo que sugiere un rol significativo para la dinamica molecular [62]. Se postula que un aspecto de la dinamica molecular relevante para la diversificacion de especificidad por sustrato es el numero de conformeros [63]. Por ejemplo, en la actinobacteria *Corynebacterium diphtheriae* parece que el contexto genomico correlaciona con perdida de promiscuidad de PriA ya que al poseer el genoma una copia de *trpF*, la enzima perdió esta funcion quimica conservando solo la funcion EC 5.3.1.16 correspondiente a la ruta de histidina. Esta sub-funcionalizacion se refleja en la perdida de estados conformacionales cambiando desde 1 estado en *C. diphtheriae* hasta 4 presentes en la dinamica de PriA de *M. tuberculosis* [51].

Las regiones rígidas de una enzima proporcionan orientacion adecuada con respecto a los grupos cataliticos, mientras que las regiones flexibles permiten al sitio activo adaptarse a los sustratos con diferentes formas y tamaños [5]. Esta consideracion sugiere que la flexibilidad del sitio activo es otra caracteristica de la dinamica molecular a considerar para obtener informacion de la capacidad de ligacion de una enzima a distintos sustratos [64]. Recientemente el indice de flexibilidad dinamica (dfi) se utilizo como una medida cuantitativa basado en la respuesta a perturbaciones de aminoacidos (PRS). Este indice se incremento en regiones cercanas al sitio activo de beta lactamasas promiscuas respecto al correspondiente dfi de  $\beta$  -lactamasas especialistas existentes [16].

### 0.1.6 Modelo biologico diversidad de Actinobacteria

Al escoger un conjunto acotado para investigar familias de enzimas promiscuas se debe recordar que la funcionalidad es jerarquica por lo que para mejorar la anotacion, es deseable reflejar el proceso evolutivo y restringirse a un grupo de organismos taxonomicamente relacionados [65]. Actinobacteria es un phylum que posee promiscuidad tanto en el metabolismo periferico como en el core metabolico. Entre datos publicos (NCBI) y privados estan disponibles alrededor de 1200 genomas no redundantes de especies de Actinobacteria. Como punto de partida, se han estudiado las relaciones filogeneticas y grupos de ortologia [66], en particular en Actinobacteria para identificar relaciones entre las familias del phylum, se obtuvieron arboles multilocus de entre 100 y 157 genomas [68]. Estos estudios sugieren como separar los genomas disponibles para hacer el calculo de grupos de ortologia. Finalmente, se han realizado estudios de plasticidad genomica en *Streptomyces* considerando 5 y 17 organismos de los 300

genomas disponibles en la actualidad [70] donde reportan 2,018 familias en el core genome y 32,574 en el pangenoma.

### **0.1.7 Modelo metabolico biosintesis de aminoacidos.**

Al hacer el calculo vemos que Streptomyces, un genero del phylum Actinobacteria cuenta en su genoma con un promedio de 8316 secuencias codificantes segun la especie. Gran parte de estas secuencias pueden ser agrupadas en subsistemas metabolicos como metabolismo de carbohidratos o de lipidos; de estos subsistemas uno de los mas amplios es el metabolismo de aminoacidos con entre 429 y 910 secuencias segun el organismo. La sintesis de aminoacidos es un subsistema presente en todas las especies pero con suficientes variaciones que permiten hacer observaciones evolutivas. En un gran numero de Actinobacterias las rutas de histidina y triptofano de 7 y 11 pasos respectivamente convergen en una enzima bifuncional llamada PriA, que realiza tanto la funcion de HisA como la de TrpF [8]. La cantidad de familias en el subsistema de metabolismo de aminoacidos, su variabilidad, su conservacion entre distintos grupos taxonomicos y la existencia de estos ejemplos en Actinobacteria lo posicionan como un buen punto de partida para la busqueda de promiscuidad tanto de familias promiscuas como de miembros promiscuos de las mismas.

## **0.2 Antecedentes**

En las cuatro decadas de estudio de la promiscuidad enzimatica, hemos aprendido que es un fenomeno distribuido en distintos subsistemas metabolicos [72] y que su existencia puede deberse tanto al desarrollo de nuevas funciones para fines adaptativos [20], como al rescate de una funcion perdida [30]. Por ello la dinamica de perdida y ganancia de genes asociada al contexto genomico en bacterias se relaciona con cambio en la funcion enzimatica [44]. Precisando, respecto a la ganancia de genes, se postula que la bifuncionalidad precede la duplicacion [40]. Lo que implica que dada una duplicacion muy posiblemente previamente la promiscuidad estuvo presente [74].

Se han desarrollado tecnicas bioquimicas y metabolicas de medicion [15], asi como algoritmos computacionales de predicción de promiscuidad [52]. Un aspecto a mejorar dentro del modelado es la restriccion del conjunto de estudio a un grupo taxonomico tan reducido que exista congruencia en las familias de ortologia y a la vez tan amplio que permita observar efectos evolutivos; el phylum Actinobacteria ha probado tener ejemplos de promiscuidad. Si bien la secuencia no ha sido suficiente para la correcta predicción de promiscuidad [45], es posible que dentro de las tecnicas computacionales la flexibilidad durante la dinamica molecular este correlacionada con la promiscuidad de los miembros de una familia [61].

### 0.2.1 Modelo biológico

De los mas de mil genomas actualmente disponibles de Actinobacterias, se seleccionaron 888 (correspondientes a 49 familias), que no estan excesivamente fragmentados; es decir con un estimado de al menos 5 genes por contig (Tabla 2). Estos genomas fueron divididos en tres grupos ([http://pubseed.theseed.org/wc.cgi?request=show\\_otus&base=/homes/nselem/Data/CS](http://pubseed.theseed.org/wc.cgi?request=show_otus&base=/homes/nselem/Data/CS)), uno de ellos correspondiente a Streptomycetaceae, la familia con la mayor cantidad de genomas disponibles; los otros dos grupos siguieron la taxonomia propuesta por Gao & Gupta en 2012. En el grupo de 290 genomas de Streptomycetaceae 2,126,832 ORFS fueron clasificados en 288,390 familias; de las 919,292 ORF del grupo I de Actinobacteria resultaron 269,406 familias. Las relaciones taxonomicas fueron corroboradas con algoritmos propios basados en best bidirectional hits (BBH).

### 0.2.2 Subsistemas metabólicos

Los operones his y trp de histidina [76] y triptofano [77] respectivamente, participantes del metabolismo de aminoacidos estan ampliamente distribuidos en los organismos bacterianos. En Actinobacteria la familia promiscua PriA participa en ambas rutas biosinteticas, para su estudio se han generado datos bioquimicos, genomicos y estructurales (Tabla 3). En bacterias gram negativas estan presentes los operones his y trp y en lugar de PriA su familia homologa HisA. PriA comprende un conjunto de subfamilias en Actinobacteria. En Streptomyces, el gen trpF se desplaza de la vecindad genomica de trp, con lo que el homologo de hisA gana promiscuidad aunque con baja actividad de TrpF, a esta subfamilia se le llama PriB [78]. En otras Actinobacterias trpF se pierde totalmente y la familia homologa de HisA, se vuelve promiscua [8] realizando tanto la funcion quimica correspondiente a HisA como la de TrpF. Finalmente en la familia subHisA se pierde la funcion TrpF debido posiblemente a la ganancia del operon trp completo [51] y en la familia subtrpF se conserva solo a la funcion TrpF debido a la perdida del operon his [Juarez vazquez et al 2015 in prep]. Existen al menos 43 familias de Actinobacteria sin explorar respecto a la funcionalidad de PriA.

### 0.2.3 Contexto y vecindades genomicas

En 2012 fueron analizados 102 genomas de 29 familias de Actinobacteria [79]. sugiriendo que al menos en Corynebacteria el contexto y la vecindad genomica incidian en la sub-funcionalizacion de PriA en subHisA [51]. Respecto a IlvC, otra familia involucrada en la sintesis de aminoacidos fue estudiada y caracterizada bioquimicamente en 1 Corynebacterium y 8 Streptomyces [45]. Para ampliar estos resultados, utilizando la anotacion de RAST y una generalizacion de la definicion de vecindad de STRING, se diseño un algoritmo para identificar vecindades similares asi como

uno de visualizacion de contexto, ambos disponibles como software libre en [github nselem/perlas](#) .

El algoritmo de clasificacion de vecindades permite agruparlas en clusters y calificar estos clusters segun su conservacion dado un grupo de bacterias. La definicion de vecindad y similitud de vecindad esta descrita posteriormente en los metodos. El algoritmo fue aplicado a la familia IlvC en 290 Streptomyces resultando 9 clusters Datos entre los mas poblados el primero cuenta con 279 elementos, otro con 9 elemento y dos mas con 7 miembros (Fig 3), resultados experimentals son congruentes con que existe divergencia funcional entre miembros de clusters distintos [45]

#### **0.2.4 Metodos bioinformaticos**

Al evaluar PROMISE [52] en un set de datos de la familia HisA/PriA [55] obtuve que su mejor desempeño es con huella molecular de tamaño 6, donde clasifica correctamente casi todas las no promiscuas, (HisA) pero no sucede lo mismo con la familia PriA donde tiene exito en 16 de 45 casos. Al aplicar el mismo tamaño de huella a 9 miembros promiscuos de la familia IlvC no consigue predecir correctamente ninguno de ellos. Por lo menos para estas familias el conjunto de entrenamiento o los descriptores no son suficientes para la anotacion de promiscuidad.

#### **0.2.5 Evomining**

Evomining es una plataforma bioinformatica pensada para la identificacion de productos naturales que tiene entre sus exitos la identificacion de la biosintesis de arsenolipidos [65]. La busqueda de productos naturales cuenta entre sus premisas que estos se producen en vecindades genomicas llamadas clusters y que ademas clusters cercanos (ya sea en contenido genico o en la secuencia de sus componentes), exploran variaciones metabolicas, es decir sus enzimas catalizan reacciones sobre sustratos parecidos aunque no identicos [65]. La base de evomining es que las enzimas de metabolismo secundario son expansiones distantes de enzimas de rutas centrales, lo que da idea de la quimica que realizan dichas expansiones dejando por identificar el sustrato sobre el que trabajan. La primera version de evomining cuenta con 200 genomas de Actinobacteria, una base de datos de secuencias de enzimas de productos naturales y otra base de datos de secuencias de enzimas de rutas centrales curada a mano. Evomining esta ligada con el problema de la promiscuidad porque en estas familias expandidas ya sea por duplicacion o por transferencia horizontal, las expansiones pueden retener la funcion quimica de las rutas centrales y viceversa, la funcion quimica expandida suele estar presente antes de la duplicacion.

Si se combinara evomining con la premisa de que vecindades distintas son marcadoras de funciones quimicas distintas, al encontrar una familia expandida con vecindades genomicas diferentes se podria solventar la deficiencia de otros metodos bioinformaticos consistente en que para identificar familias promiscuas se debe conocer previamente

un miembro promiscuo de la misma. (Fig 4) Asi pues al combinar evomining con herramientas de vecindad genomica tanto de comparacion como de visualizacion estaremos mejorando su funcionalidad en la identificacion de familias promiscuas.

### 0.2.6 Caracterizacion in vivo

Algunas enzimas PriA no han mostrado promiscuidad in vitro pero si in vivo ya que sobreviven en un medio sin triptofano, es decir in vivo complementan la funcion trpF. Para la construccion de cepas de Streptomyces con variantes no nativas de priA minimizando la modificacion genomica y el efecto de sobreexpresion, se planea utilizar E. coli como intermediario para realizar seleccion por auxotrofia. Se cuenta con un conjunto de plasmidos para transformar a E. coli asi como con las mutantes sencillas de E. coli para trpF y hisA que permiten realizar seleccion por auxotrofias. Ademas tenemos una colección de cepas nativas de Streptomyces asi como un mutante de PriA de S. coelicolor. Se optimizo una reaccion de PCR para la amplificacion de un segmento de DNA de S. coelicolor que contiene a priA.

### 0.2.7 Caracterizacion bioquimica in vitro.

De la familia PriA y sus subfamilias se han caracterizado bioquimicamente miembros selectos de Actinomycetaceae, Bifidobacteriaceae, Micrococcaceae, Acidimicrobiaceae, Corynebacterium, Mycobacteriaceae, Streptomycetaceae, Camera (provenientes de metagenoma), reconstrucciones ancestrales, 80 mutantes de Corynebacterium, y 2 mutantes de Camera mediante cineticas enzimaticas para calcular las constantes Kcat,Km. El genero Streptomyces, el que cuenta con mayor cantidad de genomas disponibles representa una oportunidad muy poco explotada de explorar la influencia del contexto y la vecindad genomicas en secuencias de PriA (Tabla 3, Figura 5).

### 0.2.8 Modelado de dinamica molecular

La dinamica es un metodo que permite hacer simulaciones de particulas que sirve para obtener informacion de propiedades macroscopicas de un conjunto de atomos [80]. Es util en el marco de mi proyecto porque permite la exploracion del espacio conformacional, y se ha visto que este esta relacionado con la actividad de la enzima [82], ademas dado un conformero permite verificar su estabilidad. Resuelve la ecuacion de movimiento de Newton con base a una configuracion inicial, las fuerzas interatomicas como los enlaces covalentes, las fuerzas de Van der Waals y la carga de las particulas[58]. Entonces para generar una simulacion de dinamica molecular, debe contarse con una estructura como punto de partida, ya sea esta cristalografica o modelada de novo o por homologia. El laboratorio de bioinformatica y biofisica computacional ha desarrollado un protocolo de generacion de modelos homologos estructurales y dinamicas moleculares (Carrillo-Tripp et al 2015 in prep); con este pipeline se han

generado dos estructuras de Camera [55], 30 estructuras y dinamicas de miembros de Actinobacteriaceae y Bifidobacteriaceae (Vazquez-Juarez et al in prep.) y finalmente una estructura de subHisA de *Corynebacterium diphtheriae*. En la familia Streptomyces, interesante debido a su variacion en contexto genomico y en mediciones in vitro aun no se modelan dinamicas moleculares aunque 40 estructuras por homologia estan en proceso.

En un estudio de subHisA [79] se utilizo el metodo de dinamica molecular y se comparo el numero de conformeros entre miembros de subHisA y PriA, resultando mayor el de PriA como corresponde a una enzima promiscua. El estudio sobre la relacion dinamica-flexibilidad de  $\beta$ -lactamasas utiliza replica exchange, una variacion de dinamica molecular que corre replicas en paralelo a distintas temperaturas [83]. Una desventaja de este metodo es que por el costo computacional de las replicas agregar explicitamente otras moleculas a la simulacion como el solvente no es posible en tiempo razonable. Una vez generadas las dinamicas moleculares se procedera a calcular tanto el numero de conformeros como el indice de flexibilidad dsi [16]. Se esta desarrollando PEDB, promiscuous enzyme database, una base datos genomicos, evolutivos, bioquimicos y estructurales y de metabolismo de PriA en Actinobacteria donde se procedera al analisis de los mismos (<http://148.247.230.43/nselem/PHP/queries.html>).

En conclusion la promiscuidad enzimatica es un fenomeno complejo debido a multiples causas. Existe una gran variedad de estudios con enfoques puntuales sobre aspectos estructurales, dinamicos y evolutivos de familias de enzimas promiscuas, sin embargo hasta ahora no se han reportado trabajos multidisciplinarios que involucren a todas las partes involucradas (Fig 6)

### 0.3 Objetivo General

Estudiar el fenomeno de promiscuidad enzimatica tanto desarrollando estrategias para identificar familias promiscuas dentro de un grupo taxonomico, como comparando variaciones de promiscuidad in vitro e in vivo con variaciones en contexto genomico y flexibilidad en miembros de una familia. (Figura 7)

### 0.4 Objetivos particulares

Mejorar evomining como metodo de identificacion de familias enzimaticas promiscuas aprovechando los cambios en vecindades genomicas como caracteristicas informativas provenientes de datos filogenomicos. Estudiar la relacion entre historias filogenomicas y procesos biofisicos con la promiscuidad in vitro, a traves de mediciones de ciertas caracteristicas de la familia PriA. Caracterizar cambios de promiscuidad enzimatica in vivo mediante perfiles metabolomicos de actividades de PriA y enzimas asociadas.

## 0.5 Estrategias

### 0.5.1 La promiscuidad en familias enzimaticas.

Mejorar Evomining mediante la identificacion de cambios de vecindad genomica en familias selectas de metabolismo central convirtiendola en una plataforma de codigo libre disponible para otros investigadores.

#### Obtener informacion genomica del phylum Actinobacteria.

Colectar genomas de Actinobacteria de NCBI y de colecciones privadas.

#### Anotar consistentemente las secuencias codificantes de estos genomas.

Utilizar un anotador automatizado y desarrollar los scripts necesarios para anotar los genomas.

#### Establecer las relaciones filogeneticas de los genomas colectados.

Mediante el uso del core genome construir un arbol filogenomico que permita establecer un marco sobre el cual hablar de cambio y que facilite reclasificar los genomas mal nombrados.

##### Identificar cambios en la vecindad genomica en familias selectas de enzimas de metabolismo central. Clasificar sistematicamente las secuencias de familias codificantes segun su similitud en familias enzimaticas.

Desarrollar las herramientas bioinformaticas necesarias para separar clusters de vecindades genomicas.

##### Sistematizar Evomining para convertirla una plataforma descargable y utilizable en cualquier set de datos bacterianos relacionados taxonomicamente proporcionados por el usuario.

Ampliar el contenido de Evomining al integrar los genomas colectados de Actinobacteria. Sistematizar la base de datos de metabolismo central.

Desarrollar la visualizacion e integrar la clasificacion de vecindades genomicas como una herramienta adicional en la busqueda de promiscuidad.

#### Promiscuidad in vitro dentro de miembros de una familia promiscua de enzimas.

Dados los sustratos conocidos de PriA investigar las posibles correlaciones entre mediciones de constantes cataliticas, contexto genomico, vecindad genomica, numero de conformeros e indice de flexibilidad.

#### Seleccionar miembros homologos de la familia de enzimas.

Se escogieron 41 Streptomyces repartidos en un arbol de rpoB de 400 Streptomyces con genoma disponible. Esta seleccion incluye los seis Streptomyces de los que se

cuenta con cinetica enzimatica de PriA, tres de ellos con estructura cristalografica.  
####Medir cineticas enzimaticas, contexto genomico, vecindad genomica, flexibilidad y numero de conformeros.

Determinar la pertenencia a uno de cuatro posibles contextos genomicos respecto al gen trpF. Estudiar la existencia de distintas vecindades genomicas. Determinar la cinetica enzimatica de 9 enzimas mas buscando variabilidad en contexto genomico (sugeridas en la tabla 4). Obtener mediante una colaboracion 37 modelos estructurales por homologia y modelar dinamica molecular.

La siguiente tabla contiene la diversidad de contextos y vecindades genomicas de 41 Streptomyces respecto al gen trpF.

#### **Determinar posibles correlaciones entre los datos producidos.**

Numero de conformeros e indice de promiscuidad.

indice de flexibilidad y numero de conformeros.

Numero de conformeros y contexto genomico.

indice de flexibilidad y contexto genomico.

Contexto genomico e indice de promiscuidad I.

Analizar las vecindades genomicas e indice de promiscuidad I.

#### **0.5.2 Desarrollar una metodologia para la deteccion in vivo de promiscuidad enzimatica.**

Debido a cambios en flexibilidad o cambios de contexto genomico, se puede sospechar de diferencias en la funcion quimica de dos miembros de una familia de enzimas, sin conocer las diferencias a nivel de sustratos. Para investigar estos cambio in vivo se propone estudiar diferencias en perfiles metabolomicos de una colección de cepas en condiciones diversas.

#### **Crear cepas geneticamente modificadas con variantes funcionales no nativas de PriA y enzimas asociadas.**

Dado un organismo modelo sustituir su homologo nativo de priA por una variante no nativa ya sea de priA o trpF de Actinobacterias selectas de las que se sospecha cambio en promiscuidad.

####Separar posibles productos y minimizar los falsos positivos debidos a perturbaciones metabolicas no relacionadas a PriA.

Separar los metabolitos mediante cromatografia dirigida por tamaño. Obtener un espectro de masas antes y despues de la sustitucion de la variante y sobre las diferencias en el espectro realizar espectrometria de masas en tandem (MS/MS) es decir refragmentar y analizar que los fragmentos contengan partes parecidas a los sustratos conocidos.

## 0.6 Metodologia

A continuacion describire la metodologia para cada una de las estrategias expuestas previamente. Todos los scripts desarrollados fueron escritos en perl y estan disponibles en github <https://github.com/nselem/perlas>.

### 0.6.1 La promiscuidad en familias enzimaticas.

#### Actinobacteria genomic

Para obtener informacion genomica del phylum Actinobacteria mediante la colección de genomas de NCBI se revisaron todas las familias de Actinobacteria de la base genoma de NCBI y se seleccionaron los genomas con minimo 5 genes por contig. Se crearon scripts para utilizar la interfaz e-utils de NCBI y descargar estos genomas desde la terminal a partir de una lista de identificadores.

#### Annotation

Para anotar consistentemente las secuencias codificantes de estos genomas se utilizo el anotador automatizado RAST y se desarrollaron los scripts necesarios para anotar los genomas desde la terminal, conectado asi NCBI y RAST.

#### Genomic DB phylogeny

Establecer las relaciones filogeneticas de los genomas colectados. Mediante el uso del core genome para construir un arbol filogenomico, para reclasificar los genomas mal nombrados.

Para obtener el core genome y en base a el reclasificar los genomas se diseño el algoritmo estrellas basado en Best Bidirectional Hits (blast all vs all).

Estrellas. Se realiza un blast all vs all de genomas deseados. Para cada secuencia, centrado en cada genoma se realiza una lista (estrella) de sus mejores hits bidireccionales. Si las listas de todos los genomas coinciden es un BBH multiple y se agrega la lista al core genome. (Fig 9) Una vez con el core genome completo se puede reconstruir la filogenia. Este metodo fue exitoso en la detección de una familia marcadora de *Clavibacter michiganensis* (2014 Rodriguez-Orduña in prep).

### 0.6.2 Identificar cambios en la vecindad genomica en familias selectas de enzimas de metabolismo central.

Clasificar sistematicamente las secuencias de familias codificantes segun su similitud en familias enzimaticas.

Como se menciono en los antecedentes, se han separado 888 genomas de Actinobacteria en 3 grupos taxonomicos utilizando para la anotacion la tecnologia de subsistemas de RAST. Para la separacion en familias iso funcionales (ortologos, paralogos y expansiones) no se utilizo RAST, especificamente el script What Changed (WC) que asigna un numero a cada familia, esta herramienta esta basada en k-mers, su codigo esta disponible en github: ([https://github.com/kbase/kbseed/blob/master/service-scripts/svr\\_CS.pl](https://github.com/kbase/kbseed/blob/master/service-scripts/svr_CS.pl)). Ademas de en los tres grupos ya mencionados, tambien se realizara una clasificacion para trescientos genomas de Actinobacteria distribuidos en todas sus familias taxonomicas.

Para desarrollar las herramientas bioinformaticas necesarias para separar clusters de vecindades genomicas a continuacion se describe detalladamente como se definio vecindad genomica y la relacion implementada de similitud.

1. Un conjunto expandido es un conjunto que contiene secuencias homologas asi como sus expansiones: paralogos y transferencias horizontales. Dado un conjunto de genomas, se pueden calcular y enumerar todos sus contextos extendidos utilizando WC.
2. Un PEG es un elemento de un conjunto expandido. Dado un PEG p, se define  $CE(p)$  el numero del conjunto expandido de p, como el numero asignado por WC al conjunto expandido a que p pertenece.
3. La vecindad de un PEG es el conjunto de PEGs cercanos a el. Dado un umbral en terminos de distancia de pares de bases entre puntos medios para precisar la definicion de cercano, se pueden calcular todos los contextos de un genoma.
4. Una vecindad A es n-similar a otra vecindad B si  $C = \{aA \mid bB, CE(b) = CE(a)\}$  tiene al menos cardinalidad n. Es decir si existen al menos n elementos de A que pertenecen al mismo conjunto expandido que algun elemento de B.
5. Un conjunto de vecindades es un conjunto de PEGs clusterizado segun la relacion n-similaridad. Si A es n-similar a B y B es n-similar a C entonces, aun si A no fuese n-similar a C, los PEGs generadores de A,B,C son agrupados dentro del mismoconjunto de vecindades.
6. Un cluster es un conjunto de conjunto de vecindades.
7. Los clusters son evaluados segun el numero y la cardinalidad de sus conjuntos de vecindades.

Sea  $Cl$  un cluster, donde  $CC_i$  es un conjunto de contextos y  $n_i$  es la cardinalidad de  $CC_i$   $Cl = \{CC_1, CC_2, \dots, CC_k\}$

Sean M la cardinalidad maxima de un conjunto de contextos y m la cardinalidad maxima sin considerar M.

$\#\$ \$ \quad M \neq \max\{ni\} \quad i \neq \{1, 2, \dots, k\}$

$m \neq \max ni \quad i \neq \{1, 2, \dots, k\} \quad ni \neq M$

$$\sum_{j=1}^n (\delta\theta_j)^2 \leq \frac{\beta_i^2}{\delta_i^2 + \rho_i^2} \left[ 2\rho_i^2 + \frac{\delta_i^2 \beta_i^2}{\delta_i^2 + \rho_i^2} \right] \equiv \omega_i^2$$

M representa el contexto mas difundido de la enzima, dentro del grupo taxonomico considerado; mes relevante porque si m es grande significa que hay un segundo contexto genomico conservado en dicho grupo taxonomico, y entonces posiblemente una ganancia de funcion.

La evaluacion de Cl esta dada por una combinacion lineal de k,m y M  
 $S(Cl) = f(k, m, M) = c_1k + c_2m + c_3M$

Este algoritmo se puede mejorar considerando la orientacion de los genes del cluster asi como clusters de los vecinos.

### Organizar y presentar los datos en una plataforma.

Para contribuir al desarrollo de la plataforma Evomining se desarrollaran scripts de visualizacion de arboles filogeneticos y contextos genomicos.

Para facilitar el analisis visual de una vecindad genomica ya la vez generar imagenes de alta calidad facilmente exportables para su uso en publicaciones, se desarrollaran scripts de visualizacion que utilizaran el formato Scalable Vector Graphics (SVG), dicho formato es basicamente un archivo de texto XML que contiene instrucciones para que el navegador realice un dibujo (W3school/SVG 2015). Al ser vectores, las imagenes generadas en SVG no pierden resolucion al ser escaladas y justamente por ser escalables permiten explorar con detalle grandes cantidades de datos organizados por ejemplo en arboles filogeneticos. Los scripts a desarrollar extraeran para cada gen informacion necesaria como coordenadas, direccion, funcion quimica, etc, proveniente de la anotacion de RAST y de los scripts de comparacion de vecindades genomicas. La primera version de evomining fue desarrollada en el lenguaje perl; este lenguaje cuenta con un modulo para facilitar la elaboracion de SVG (perlmaven/SVG 2015) por lo que al utilizar SVG no se agregan nuevos requerimientos a su desarrollo y se facilita su portabilidad.

Se amplificara Evomining de los 200 genomas con que contaba su version inicial a los 880 colectados mudando la curacion manual de su base de datos de rutas centrales a la anotacion por subsistemas de RAST. Finalmente se presentara la variacion en vecindades genomicos como una herramienta adicional que ayude en la busqueda de promiscuidad en familias de enzimas pertenecientes al metabolismo central.

### 0.6.3 Promiscuidad in vitro

#### Datos cineticos:

En todos los ensayos enzimaticos se busca medir una señal que permita una distincion clara entre sustrato y producto [84]. La cinetica enzimatica de PriA proveniente del genero Streptomyces sera determinada como ya se ha reportado previamente, mediante el monitoreo de cambios en fluorescencia (isomerizacion del sustrato PRA) o en absorbancia (isomerizacion del sustrato PROFAR). En el caso de la isomerizacion de PRA, debido a que contiene un anillo de antranilato, la fluorescencia del sustrato PRA es 50 veces mayor que la del producto 1-(2-carboxyphenylamino)-1-deoxy-D-ribulose 5-phosphate (CdRP) por lo que se utiliza la disminucion en fluorescencia como medida de la conversion del sustrato en producto [85]. Se mandaran sintetizar estas variantes para posteriormente sobre expresarlas en E. coli. Se creceran cepas modificadas de E. coli (W-, H-) en medio minimo M9 enriquecido con una mezcla de aminoacidos excepto L-histidina y L-triptofano y se seleccionaran por rescate de auxotrofia. Para obtener la enzima necesaria para los ensayos enzimaticos se utilizaran plasmidos disponibles para construcciones de sobreexpresion de proteina, despues de la produccion la enzima se purificara utilizando cromatografia por afinidad a niquel [78].

Finalmente se recopilaran datos cineticos de PriA tanto privados como los publicos reportados a la fecha en la BRAunschweig ENzyme Database BRENDA [86]. Una vez colectados los datos se anotaran en PEDB, nuestra base de datos ad hoc, y se tomara como medida de promiscuidad el I-index [15] que se define como:  $I = \frac{1}{N} \sum_{i=1}^N \frac{1}{K_i} \text{cat Kim}_i / \text{cat Kim}_i$

#### Dinamica molecular

Para generar dinamicas moleculares primer lugar se recolectaran las estructuras tridimensionales de miembros de PriA de Actinobacteria. Despues se procedera a modelar por homologia las estructuras tridimensionales faltantes utilizando el pipeline del laboratorio de bioinformatica y biofisica computacional. Este pipeline utiliza el software Rosetta para el modelado para las estructuras y GROMACS Groningen Machine for Chemical Simulation, [87] para el modelado de la dinamica molecular. Esta parte del trabajo se realizara en colaboracion con el laboratorio de bioinformatica y biofisica computacional.

### 0.6.4 Promiscuidad in vivo

Se realizaran construcciones con variantes no nativas de priA y/o trpF en Streptomyces coelicolor. Para las construcciones se amplificara mediante PCR un fragmento alrededor de PriA que se insertara en un vector. Este vector recombinara en E. coli con un casete provisto de un gen marcador de resistencia a antibiotico y este gen recombinado

se pasara por conjugacion a *S. coelicolor* donde se espera que realice una doble recombinacion. El paso por *E. coli* es llevado a cabo porque *Streptomyces* no se puede transformar por electroporacion. Se seleccionaran las cepas de *Streptomyces* resistentes al antibiotico como prueba de que ya no poseen su priA nativa. Posteriormente, mediante un procedimiento analogo se sustituirá el gen marcador, por variantes no nativas de priA/trpF.

La cromatografia se refiere a un conjunto de metodos que separan y analizan mezclas de moléculas. Basicamente estos metodos se basan en diferencias en el tamaño, intercambio de iones y afinidad. [58] Posteriormente se combinan con espectrometria de masas que es una tecnica que mide el radio masa-carga de las partículas fragmentadas en iones. [58]. Los datos obtenidos de espectrometria de masas se procesaran utilizando redes moleculares, que consiste en agrupar los productos segun la similitud de sus partes. Plan: 3 replicas tecnicas, 2 replicas biologicas de 5 cepas.

### 0.6.5 Consideraciones

Falsos negativos respecto a promiscuidad estan muy extendidos en la literatura y en las bases de datos, en parte porque la mayoria de las funciones son asignadas por similitud de secuencia y dado un falso negativo el error se propaga en secuencias similares. Por otro lado es muy dificil demostrar un verdadero negativo a menos que se prueben todas las posibilidades de sustrato para la enzima. Sin embargo el espacio de sustratos puede acotarse gracias a tecnicas como el docking que esta intimamente relacionado con la dinamica molecular [58]. Limitar el espacio de sustratos puede retroalimentarse con el estudio de la promiscuidad in vivo y viceversa.

Con los metodos propuestos en este trabajo solo se podra detectar perdida o ganancia de promiscuidad entre enzimas de organismos respecto a otros miembros dentro un grupo taxonomico, no asi el estado de promiscuidad intrinseco a la enzima. Si dada una enzima no se detectan variaciones en contexto, vecindad genomica o flexibilidad dentro de un grupo taxonomico cercano, entonces no podemos decir en principio nada acerca de la promiscuidad de la variante, posiblemente es promiscua pero al mantenerse constante en todos los parametros descritos, con estos metodos no se puede sugerir promiscuidad. Es posible que al mirar en un grupo taxonomico mas amplio se detecte una neofuncionalizacion de la familia aunque tambien es posible que exista una variable z como la flexibilidad de sustrato [23] que no se este considerando y que explique o sea el mejor indicador para esta familia de promiscuidad enzimática.

Se debe considerar que si existe una correlacion vecindad genomica-promiscuidad, esta no indica causa efecto, mas bien, es plausible que la vecindad sea un amplificacion de diferencias en secuencia, a un numero igual de variaciones en secuencia la existencia de un cambio de vecindad indica un proceso mas largo y mas cambios, es una amplificacion de las marcas dejadas por transformaciones funcionales.

Si bien no se resuelve el problema de anotar promiscuidad automaticamente, este trabajo pretende aprovechar que los contextos genomicos ayudan a la identificacion de

familias promiscuas para mejorar una plataforma de productos naturales, pretende tambien una confirmacion de que los cambios en la dinamica molecular ayudan a identificar los miembros mas promiscuos hacia actividades recien adquiridas, asi como tambien ser pionero en la investigacion de promiscuidad in vivo.

- Gene cluster plants[89]
- Archaeal core [90]
- Natural products genomic era[91]
- Methanosarcina reconstruction [92]
- Archaea phylum[93]
- Prediction for possible products of promiscuous enzymes[???]
- Saxitoxin [94]
- Plants clusters [95]
- MiBIG [96]
- Metagenomics on Streptomyces [97]
- Sulfolobus reconstruction [98]
- Archaeal Natural products[99]
- Computational Pangenomics [100]
- Cuántos genes “obtenidos por EvoMining” son core/ cloud/stand alone
- Qué porcentaje de genes únicos recupera EvoMining
- Eucarya paralogs reshape gene clusters [101]
- Microbial dark mater [102]
- Archaea anaerobica carbon [103] Archaea Eucarya gap loki[104]
- Archaea and eucarya[105]
- BPGA [106] genes esenciales bacteria minima[107]
- Radical [108]
- RaxML large phylogenies [109]
- R phylogenies [110]
- Streptomyces exploradores [111]
- LUCA [112] Luchando por el reconocimiento de Archaea[113],[114] The primary kingdoms [115]
- Prediccion aRchaeas [116]
- RASt archaea [117] Book Archaea [118]
- Computational methods for bacterial and archaeal genomes [119]
- Archaeas boook [120]
- Bacterial /archaeal genome [121]
- Bacteria Archaea genome [122]
- Tree of life and HGC [123]
- Genomas retrospectiva 20 años [124]
- GC content plasmido genoma [125] Genoma minimo[126]
- Phylogeny R [127]
- Cyanobacteria fluctuacion genomica y adaptacion [128]
- Ecology of cyanobacterua [129]
- Histidine biosynthesis[130]
- PriA reconstruction [131]

Escala temporal bacterias [132]

Pangenome size [133]

variabilidad del 16s [134]

```
# 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)
```



# **Chapter 1**

## **EvoMining**

### **1.1 Introduction**

Enzyme promiscuity on metabolic families, can be looked on enzymes that are over a divergent process.

### **1.2 Gen families expansions on genomes**

#### **1.2.1 Pangenomes**

Expansions are located on pangenome, Tools to analyse pangenome BPgA

### **1.3 EvoMining**

EvoMining looks expansions on prokaryotic pangenome.  
Biological idea.

EvoMining was available as a consult website with 230 members of the Actinobacteria phylum as genomic data base, 226 unclassified nBCCs, and not interchangeable central database 339 queries for nine pathways, including amino acid biosynthesis, glycolysis, pentose phosphate pathway, and tricarboxylic acids cycle. [65] EvoMining was proved on Actinobacteria Arseno-lipids

## 1.4 Pangenome

The sequenced genome of an individual in some species is just a partial print of the species genetic repertoire. Individuals can gain and loss genes.

[123] 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 prokaryotes.

Maybe HGT is an selected feature, if is the case, so could be np production.

Some archaeas has open pangenome. [37]

HGT doubles mutation rate on prokaryotes. [123] Maybe HGT is an selected feature, if is the case, so could be np production.

Some archaeas has open pangenome. [37] Shell trees converge to core trees [108]

## 1.5 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 flexibility on genomic, central and natural product databases. Evomining user distribution was developed on perl on Ubuntu-14.04 but wrapped on Docker. Docker is a software containerization platform that allows repeatability regardless of the environment. Docker engine is available 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](https://github.com/EvoMining)

Github and Dockerhub can be connected 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 [48] 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 differences on a set of genomes

- *BPGA* to analize pangenome

EvoMining Dockerization was chosen to avoid future compatibility problems, for example dependencies unavailability, or incompatibility between future versions of its software components. As much as reproducible research was a concerned while developing EvoMining app, reproducibility is also important on data analysis, for that reason this document was written using R-markdown and latex template from Reed College [136]. While R-markdown allows to write and run R code and interpolate text paragraph to explain scripts and analysis.

## 1.6 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

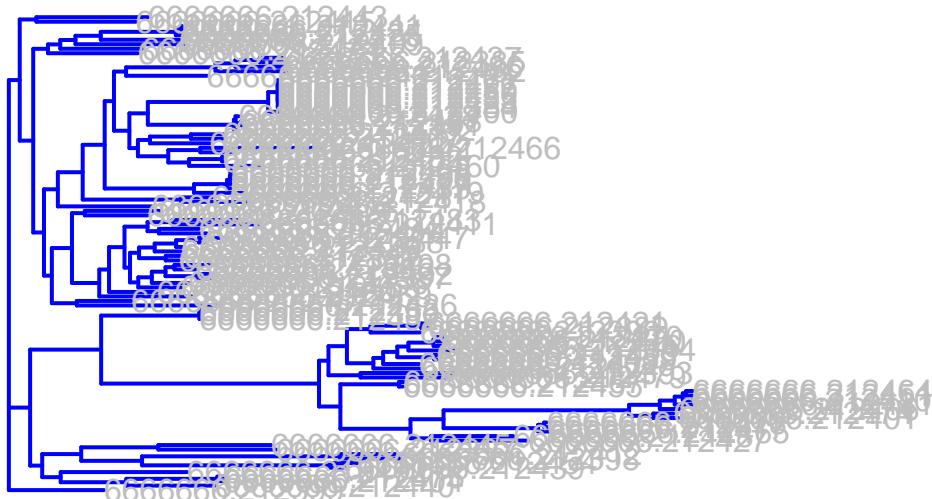
### 3. Central Pathways DB

*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 [137]

\* 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/>]

## 1.7 Data Bases

### 1.7.1 Central pathways

Central database were chosen by BBH from

```
table <- read.csv("chapter1/WC_Central/BBH_Organisms.txt", row.names = 1, sep = "\t")
kable(table,   caption = "BBH_Organisms \\label{tab:BBH_Organisms}", caption.short =
```

Table 1.1: BBH\_Organisms

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

### 1.7.2 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 organisms were

```
table <- read.csv("chapter1/WC_Central/WC_Organisms.txt", row.names = 1, sep="\t")
kable(table, caption = "WC_Organisms \label{tab:WC_Organisms}", caption.short = "WC_Org")
```

Table 1.2: 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

### 1.7.3 AntisMASH optional DB

AntiSMASH is [138]

### 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 these expansions derived from a genomic duplication?

Has Archaea natural products detected by antisMASH, and if not, where are these NP's or may Archaea doesn't have NP's.

applying EvoMining to Archaea

### 1.7.4 Otras estrategias para los clusters Argon context Idea

## 1.8 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
svrRepresentativeSequences -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 interested 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
```

### 1.8.1 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  $2\pi$  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].

## 1.9 Recomendaciones de Luis

Para evoMining

Probar distintos métodos de filogenia y después hacer la coloración.

maximum likelihood, Protest phym

Atracción de ramas largas.

raxml

trim all vs Gblocks (Tony Galvadon)

Comparar dos árboles

Para ver si la evolución de los genes concatenados ha sido simultánea

Robinson and foulds

Joe Felsenstein

Phylogenetic

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

## 1.10 CORASON: Other genome Mining tools context-based

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

Genome fluidity on Bacteria is source of biosynthetic gene clusters (BGCs) abundance, in fact almost all bacterial genome sequenced contributes with new genes and gene clusters to the Bacterial Pan-genome. As a consequence of gene diversity helped by sequence technology advances, researchers often have a large set of genomes that wish to analyze in search of a particular gene cluster variation. Answering BGCs analysis needs CORASON allows users to find and visualize variations of a given gene cluster sorting them according to the conserved core phylogeny.

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.

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

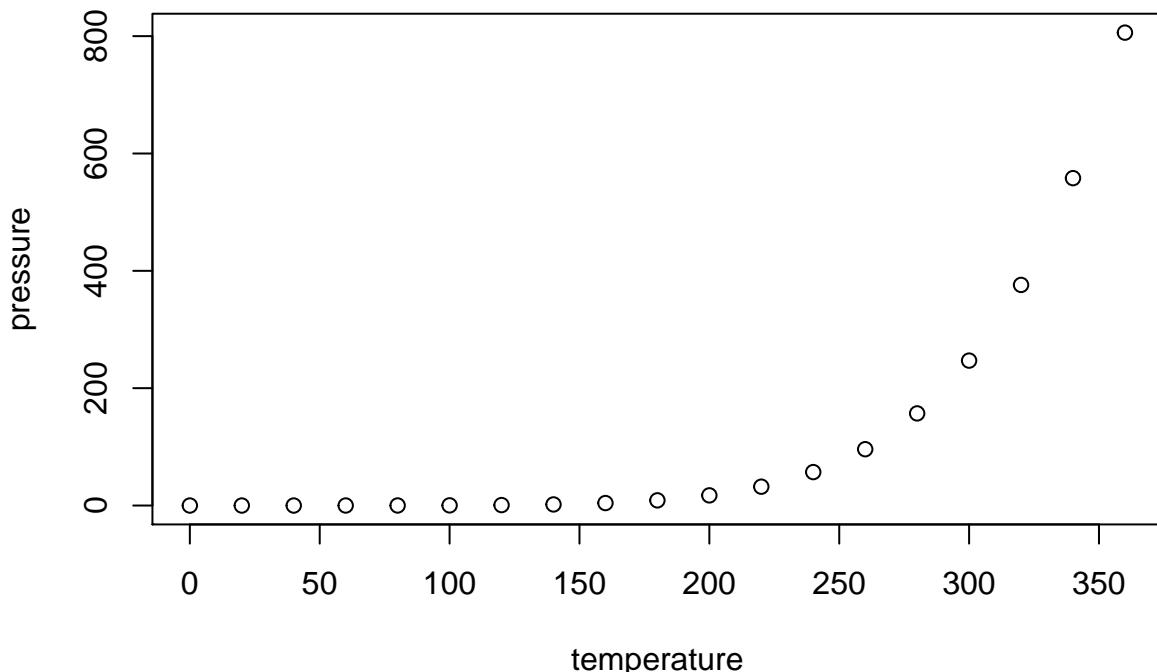
Functional annotations are provided by RAST annotation service due to that CORASON genomic databases must be RAST files. Any archaeal or bacterial genome can be RAST annotated either on the website or by command line using myrast server.

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](https://nselem.github.io/).

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 visualize variations of a reference gene cluster identifying its core functions and finally sorting variations according to their evolutionary history helping to prioritize clusters that may be involved on chemical novelty.

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.

## 1.12 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:

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 1.3: 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

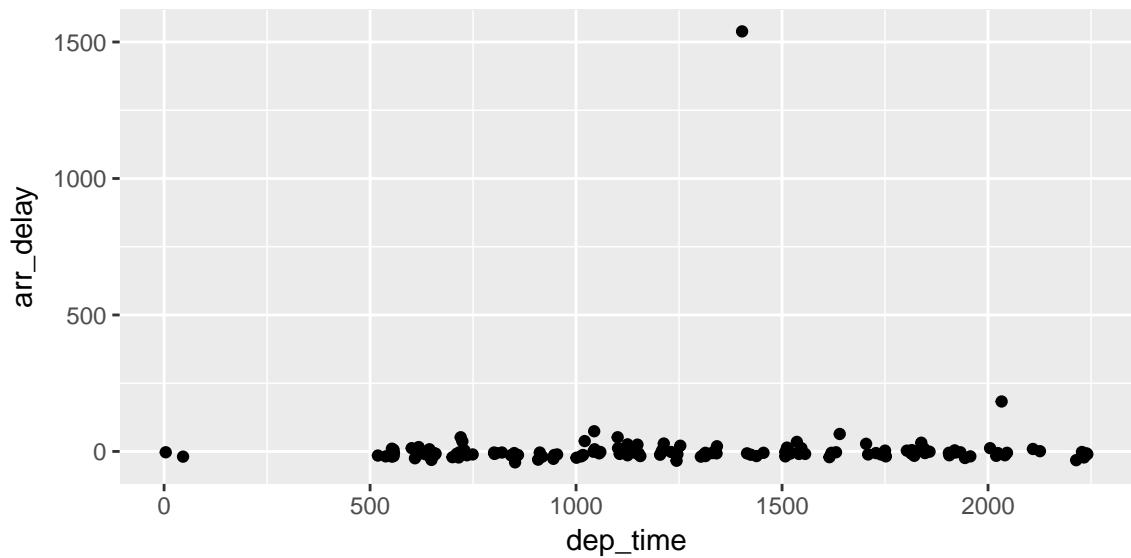
We can further look into the properties of the largest value here for American Airlines Inc. To do so, we can isolate the row corresponding to the arrival delay of 1539 minutes for American in our original `flights` dataset.

```
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()
```





# Chapter 2

## PriA Family

PriA is an homologous of HisA,  
on Actinobacteria phylum, TrpF is gene lost.

PriA is a promiscuous enzyme family involved in histidine and triptophan pathways performing HisA and TrpF functions respectively. This enzyme family catalyzes conversions converting ProFAR into PRFAR acting as a HisA and isomerizes PRA into CdRP acting as TrpF.

PriA shows a functional gradient on Actinobacteria that decide this family into enzymatic subfamilies. Between PriA subfamilies there is PriB family on Streptomyces with little trpF activity, subHisA on Corynebacteria and Actinomyces, subTrpF on Actinomyces,

## 2.1 Enzymes

39 Streptomyces PriA sequence 39 Streptomyces sequences, as outgroup E coli, Arthrobacter Aurescens, Salmonella enterica and Acidimicrobium ferrooxidans PriA's were included.

as an outgroup with HisA activity Enterobacterias *Salmonella enterica* (PDB:5AHE), *Escherichia coli K12 Acidimicrobium ferrooxydans* (PDB:4WD0) Additionally to the sequences selected by phylogeny, Jonesia denitrificans and Streptomyces sp Mg1 TrpF sequences were added as control .

PriA

*Mycobacterium tuberculosis* (Mtub PDB:2Y88,2Y89,2Y85,3ZS4) *Streptomyces coelicolor* (Scoe PDB:2VEP,2X30,1VZW),*Streptomyces globisporus*, *Actinomyces urogenitalis* 4X2R *Corynebacterium jeikeum*

subHisA

*Corynebacterium diphtheriae* *Actinomyces car* (PDB:4X2R)

subTrpF

*Arthrobacter aurescens* (PDB:4WD0)

PriB

*Streptomyces ipomoeae*, *Streptomyces sviceus* (PDB:4U28,4TX9)

TrpF controls *Jonesia denitrificans* (PDB:4WUI) *Chlamidya thrachomatis*, *Streptomyces* sp. Mg1 TrpF and *Actinomyces odontolyticus* were included

Streptomyces 34 paralogous

## 2.2 CORASON

PriA All streptomyces have a partially conserved PriA cluster. CT34 has a secondary copy whose Best hit on NCBI is Lentzea's PriA with 50% identity 98% coverage

TrpF1 TrpF1 queries gave hits with TrpC enzyme present on every Streptomyces, additionally S rimosus, S coelicolor, S venezuelae and S. NRRL S-1813 had an extra copy. S rimosus TrpC vicinity has PKS and siderophore genes.

TrpF2 Conserved cluster with NRPS sequences flanking TrpF2

TrpF3 Non conserved cluster

TrpF4 purpeofuscus and S bikiniensis 2. Heatmap

## 2.3 Structures

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

Table 2.1: Enzyme PDB

	Organismo	Family	Observations	Resolut
4WUI	Jonesia denitrificans	TrpF		1
4X9S	Streptomyces sp. MG1	PriB		1
5DN1	Streptomyces coelicolor	PriA		1
1DL3	Thermotoga maritima	TrpF		2
1LBM	Thermotoga maritima	TrpF	RCDRP	2
1NSJ	Thermotoga maritima	TrpF		2
1V5X	Thermus thermophilus	TrpF		2
1VZW	Streptomyces coelicolor	PriA		1
2VEP	Streptomyces coelicolor	PriA		1
2X30	Streptomyces coelicolor	PriA	R139N	1
2Y85	Mycobacterium tuberculosis	PriA	RCDRP	2
2Y88	Mycobacterium tuberculosis	PriA	D11N PRFAR	1
2Y89	Mycobacterium tuberculosis	PriA	D11N	2
3ZS4	Mycobacterium tuberculosis	PriA	PRFAR	1
4AAJ	Pyrococcus furiosus	TrpF		1
4TX9	Streptomyces sviceus	PriB	ProFAR	1
4U28	Streptomyces sviceus	PriB		1
4W9T	Streptomyces sp. Mg1	PriB		1
4WD0	Arthrobacter aurescens	PriB		1
4X2R	Actinomyces urogenitalis			1
4AXK	Corynebacterium efficiens	SubHisA		2
5AHE	Salmonella enterica	HisA		1
5AB3	Salmonella enterica	HisA	D7N, D10G, dup13-15, Q24L, G102A	1
5ABT	Salmonella enterica	HisA	D7N, G102A, V106M, D176A	1
5AC7	Salmonella enterica	HisA	D7N, D10G, dup13-15	1
5AC8	Salmonella enterica	HisA	D10G, dup13-15, G102A	1
5AC6	Salmonella enterica	HisA	D7N, D10G, dup13-15, Q24L, G102A	1
5A5W	Salmonella enterica	HisA	HisA D7N D176A with ProFAR	1
5AHF	Salmonella enterica	HisA	HisA D7N with ProFAR	1
4GJ1	Campylobacter jejuni	HisA		2
2W79	Thermotoga maritima	HisA		1
1QO2	Thermotoga maritima	HisA		1
5LHE	Thermococcus kodakaraensis	TrpF		1
5LHF	Thermococcus kodakaraensis	TrpF		1

```
## Substrates
```

20 substrates were collected from literature and chemoinformatics predictions.

S13,S14,!5,S16 light activated substrates, S3 PRA, S7 PROFAR, S6 GMP, S11 GTP, S17 PRAP, S18 Compound V, S17 PraP

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

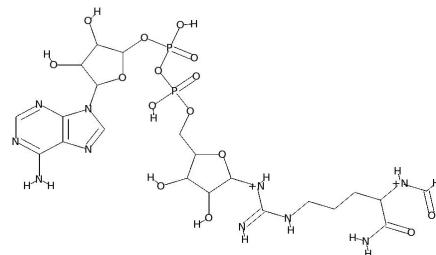
Table 2.2: Substrates

	id	Number	Kind	Reference	Names
S13	dte6_open	LUZ			
S15	dte13_open	LUZ			
S14	dte6_closed	LUZ			
S16	dte13_closed	LUZ			
S10	C04376		James		5'-Phosphoribosyl-N-formylglycinamide
S12	C03838		James		5'-Phosphoribosylglycinamide
S9	C04640		James		2-(Formamido)-N1-(5'-phosphoribosyl)acetamide
S18	CompoundV		Adams	CompoundV	This compound is an intermediary between GTP and C04640
S5	C05923		James		2,5-Diaminopyrimidine nucleoside triphosphate
S4	C05922		James		Formamidopyrimidine nucleoside triphosphate
S8	C01268		James		5-Amino-6-(5'-phosphoribosylamino)uracil
S17	PraP		Verduzco	PraP	
S7	C04302		James		PRA
S6	C00144		James		GMP
S11	C00044		James		GTP
S1	C01253		James		ADP-D-ribosyl-[dinitrogen reductase]
S2	C01201		James		N(omega)-(ADP-D-ribosyl)-L-arginine
S3	C04896				ProFAR
S19	S_17146	Due et al	17146		2,5-dimethyl-N-(4-oxocyclohexa-2,5-dienylidene)
S20	S_16827	Due et al	16827		(E)-N-(3-chloro-5-methyl-4-oxocyclohexa-2,5-dienylidene)

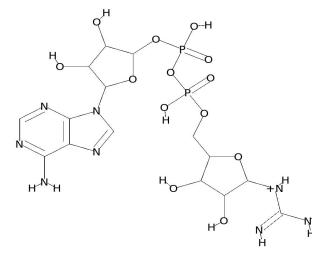
## 2.4 Docking

Docking simulation were calculated for PriA Streptomyces enzymes. TrpF enzymes from Strptomyces Mg1, Jonesia denitrificans, were added as controls

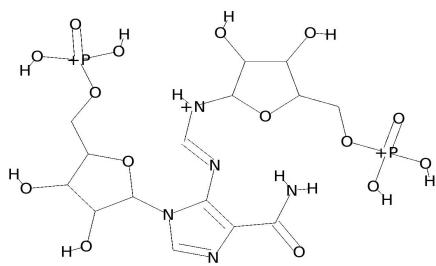
Procedures can be found at Docking Protocols



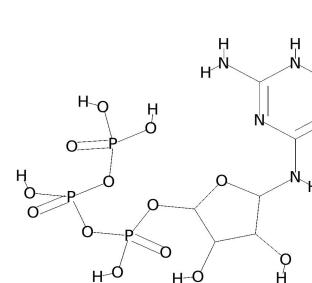
**1**  
ADP-D-ribosyl-[dinitrogen reductase]



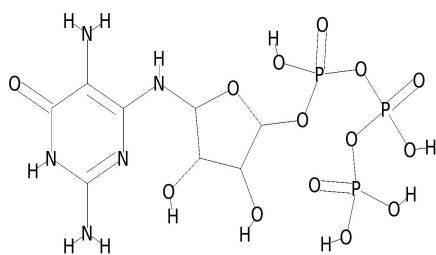
**2**  
N(omega)-(ADP-D-ribosyl)-L-amino acid



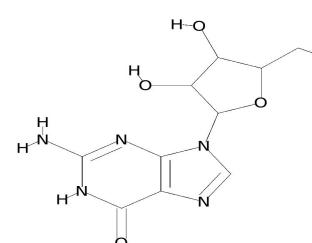
**3**  
5-(5-Phospho-D-ribosylaminoformimino)-1-(5-phosphoribosyl)-imidazole-4-carboxamide



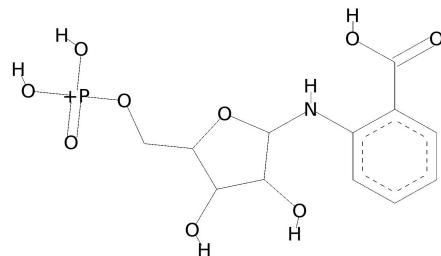
**4**  
Formamidopyrimidine nucleotide triphosphate



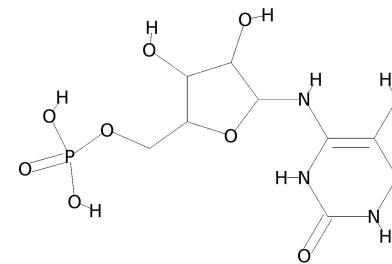
**5**  
2,5-diaminopyrimidine nucleoside triphosphate



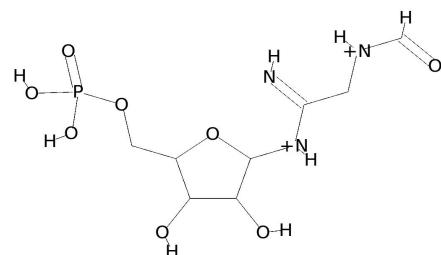
**6**  
Guanosine monophosphate



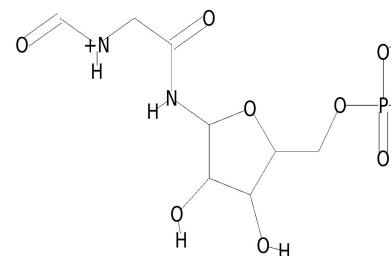
N-(5-phospho-D-ribosyl) anthranilate



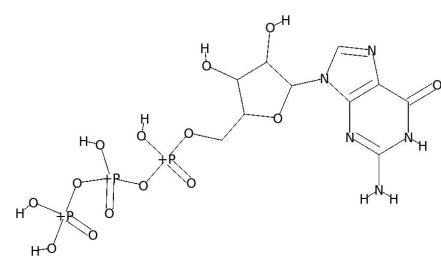
5-Amino-6-(5'-phosphoribosylamino)uracil



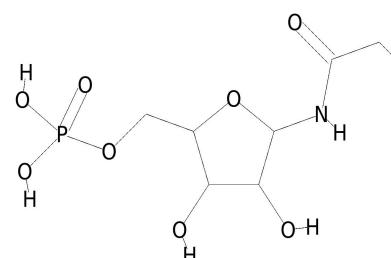
2-(Formido)-N1-(5'-phosphoribosyl)acetamidine



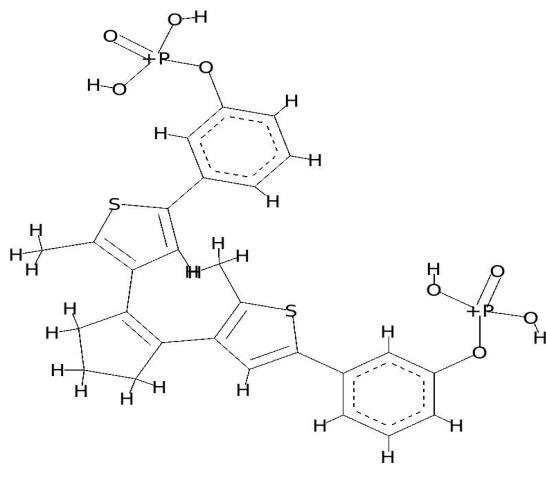
5'-Phosphoribosyl-N-formylglycinamide



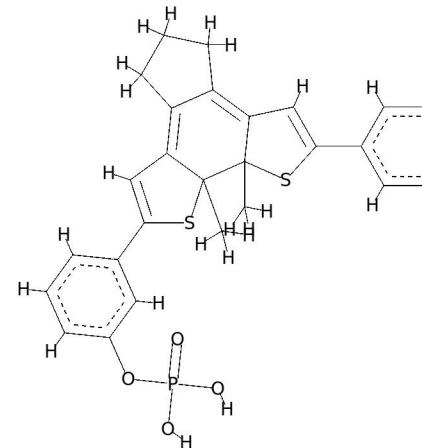
Guanosine 5'-triphosphate



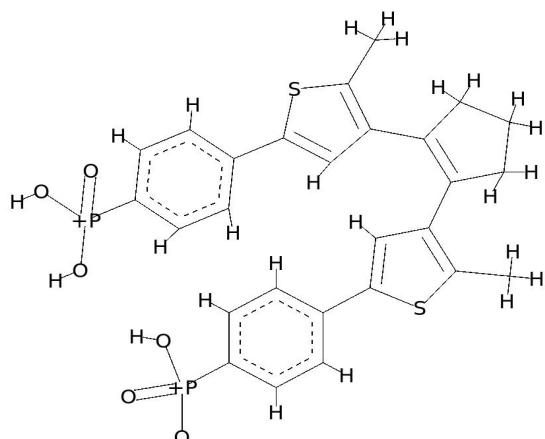
5'-Phosphoribosylglycinamide



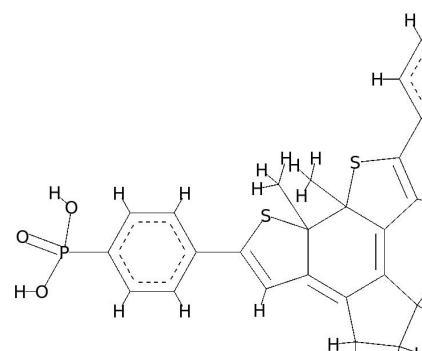
**13**  
DTE-meta-phosphate(dte6\_open form)



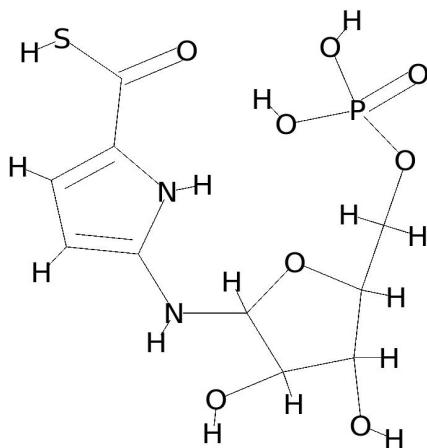
**14**  
DTE-meta-phosphate(dte6\_Closed)



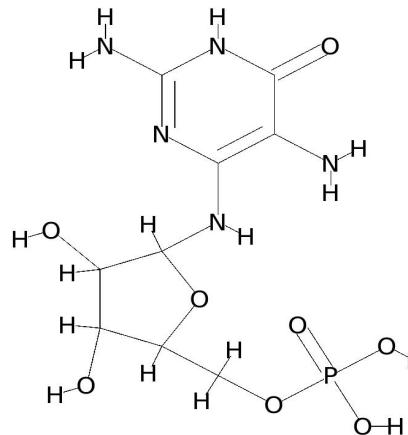
**15**  
DTE-Para-Phosphonate(dte13\_closed form)



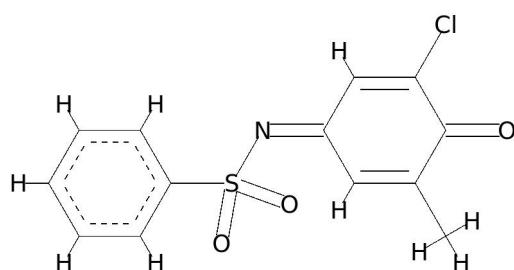
**16**  
DTE-para-phosphonate(dte13\_clo)



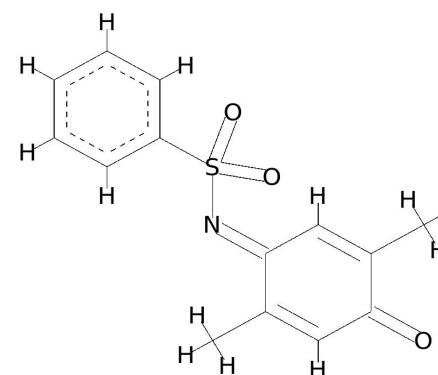
**17**  
4N'-(5'-phosphoribosyl) 4-aminopyrrole  
-2-carboxilate



**18**  
2,5-di-amino-6-ribosylamino-4-(3H)-pyrimidinone 5'-phosphate



**19**  
(E)-N-(3-chloro-5-methyl-4-oxocyclohexa-  
2,5-dienylidene)benzenesulfonamide



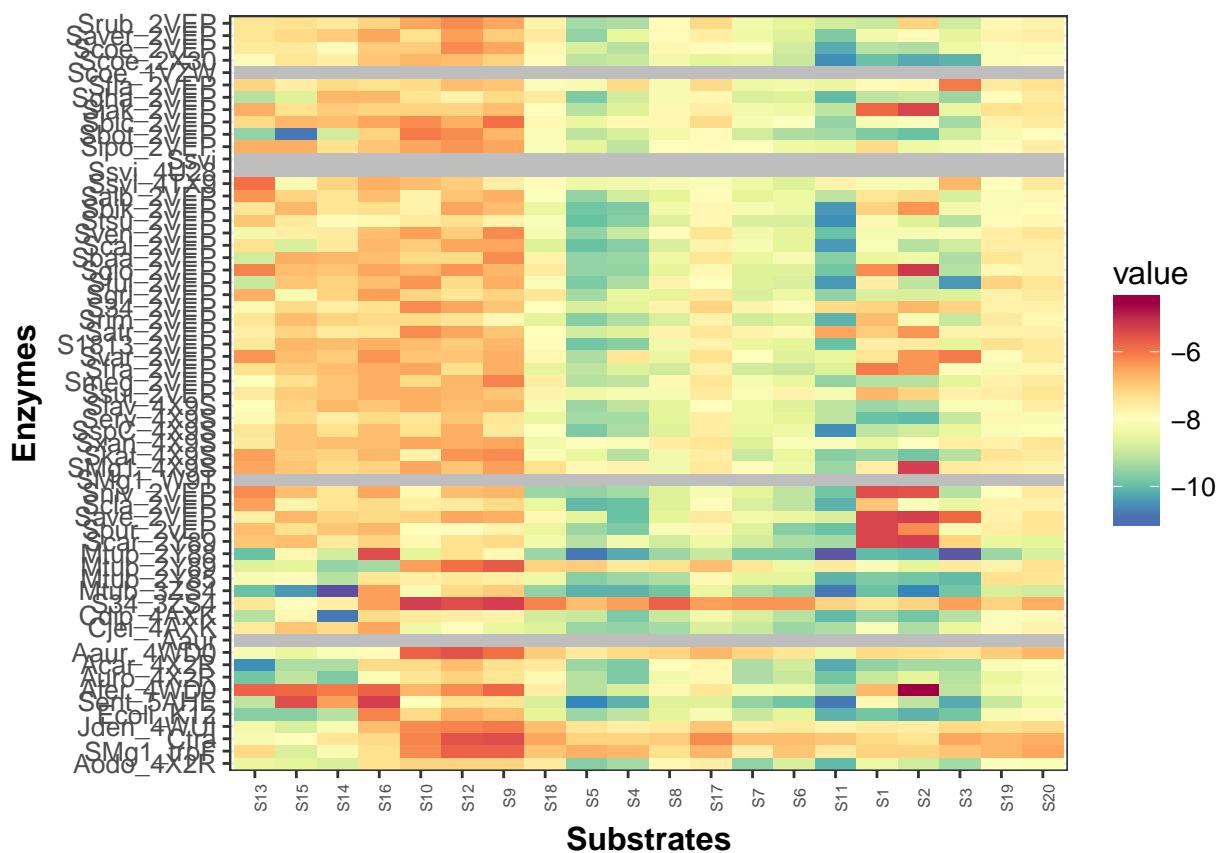
**20**  
2,5 dimethyl-N-(4-oxocyclohexa-  
2,5-dienylidene)benzenesulfonamide

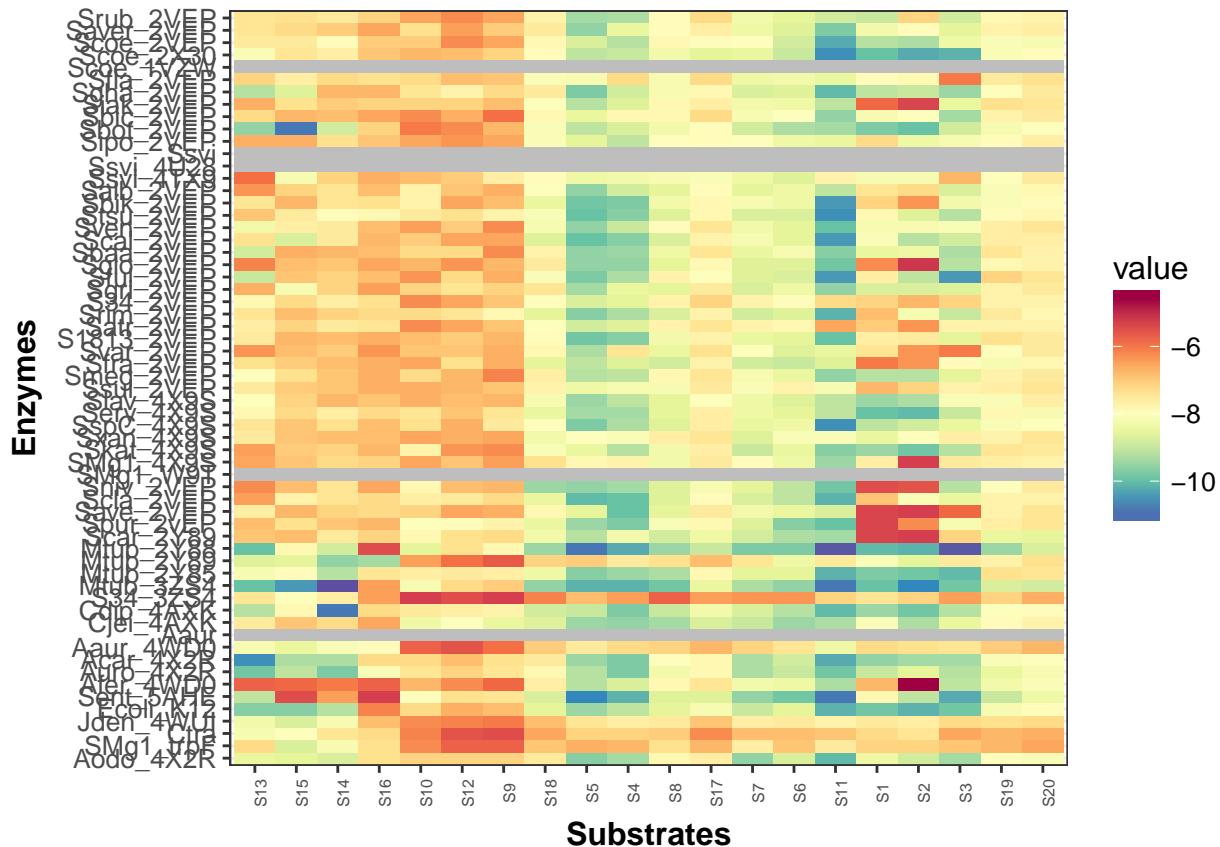
```
docking <- read.csv("chapter2/SmallHeat.data", header=TRUE, sep="\t")
kable(docking,  caption = "Enzymes docking \label{tab:docking}",caption.short = "
```

Table 2.3: Enzymes docking

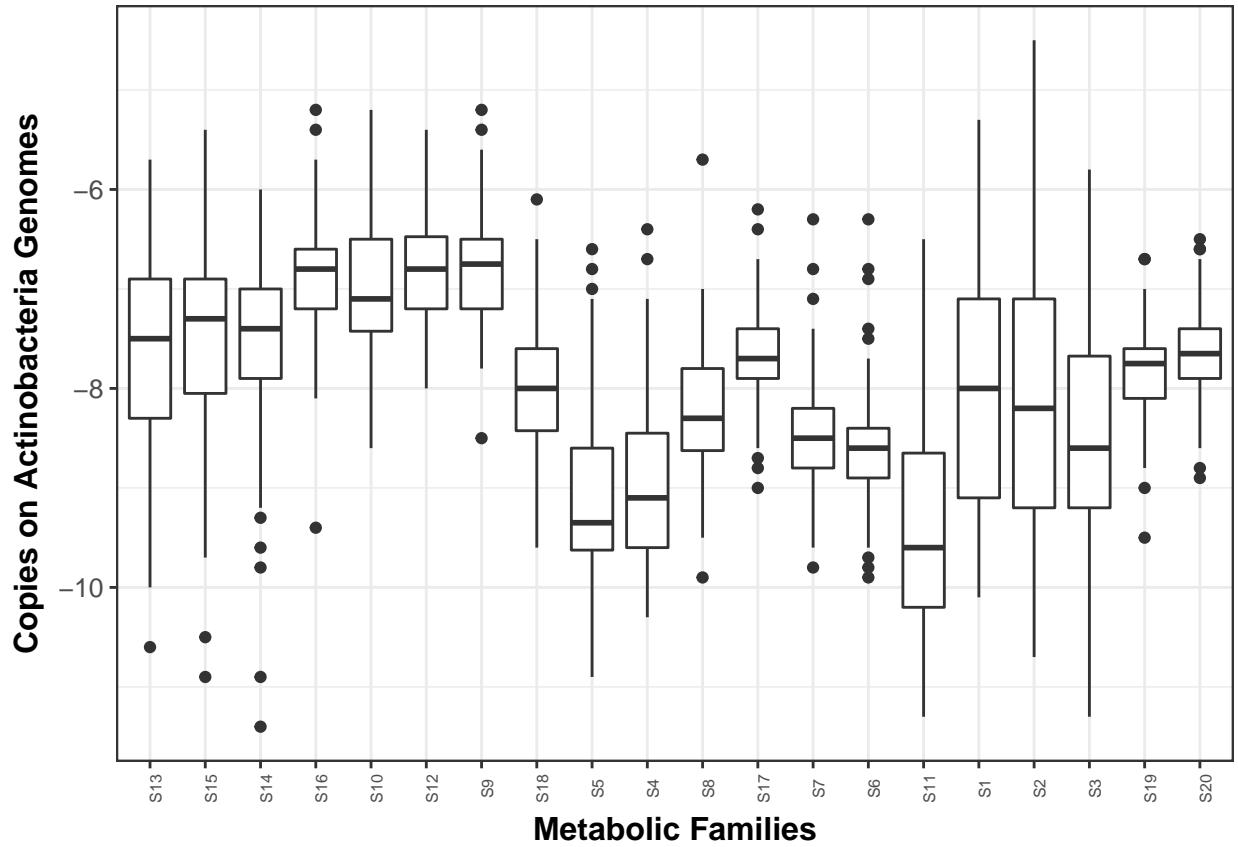
Enzima	S13	S15	S14	S16	S10	S12	S9	S18	S5	S4	S8	S17	S7	S6
Sniv_2VEP	-6.2	-6.8	-7.4	-6.5	-7.8	-6.8	-6.7	-9.5	-9.6	-9.4	-8.7	-8.2	-8.6	-9.1
Scla_2VEP	-6.4	-7.7	-7.4	-7.2	-7.6	-7.5	-7.2	-8.5	-10.1	-10.0	-8.7	-7.9	-8.8	-9.1
Save_2VEP	-7.6	-6.7	-7.1	-7.2	-7.1	-6.5	-6.6	-7.8	-8.6	-10.0	-8.6	-7.5	-8.3	-8.5
Spur_2vEP	-6.8	-7.3	-6.9	-6.7	-8.0	-7.9	-7.7	-8.5	-9.5	-9.8	-8.1	-7.8	-8.7	-9.7
Scar_2Y89	-6.9	-6.8	-7.5	-7.1	-7.8	-7.3	-7.2	-8.3	-9.2	-8.3	-8.9	-8.4	-8.9	-9.3
Mtub_2Y88	-10.0	-7.8	-8.9	-5.4	-8.6	-7.3	-7.8	-9.5	-10.9	-10.3	-9.5	-9.0	-9.8	-9.8
Mtub_2Y89	-8.7	-8.6	-9.6	-9.4	-6.4	-5.9	-5.6	-7.1	-7.0	-7.5	-7.3	-6.8	-7.4	-8.4
Mtub_2Y85	-8.2	-7.9	-9.2	-7.4	-7.6	-7.5	-7.6	-8.4	-9.7	-9.5	-9.3	-7.8	-8.6	-8.6
Mtub_3ZS4	-10.0	-10.5	-11.4	-6.4	-8.2	-7.2	-7.0	-9.6	-10.2	-10.2	-9.9	-8.5	-9.3	-9.6
S34_3ZS4	-7.4	-8.0	-7.6	-6.4	-5.2	-5.4	-5.2	-6.1	-6.8	-6.4	-5.7	-6.4	-6.3	-6.3
Cdip_4AXK	-9.2	-7.8	-10.9	-7.2	-7.5	-7.6	-7.7	-8.9	-9.0	-9.8	-9.0	-8.3	-8.8	-9.2
Cjei_4AXK	-7.5	-6.9	-7.2	-6.5	-8.4	-8.0	-8.5	-8.7	-9.5	-9.6	-9.4	-8.8	-9.0	-9.5
Aaur	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Aaur_4WD0	-8.2	-8.5	-8.1	-7.9	-5.7	-5.5	-5.9	-7.0	-7.5	-7.2	-7.1	-6.7	-7.1	-7.4
Acar_4X2R	-10.6	-9.3	-9.3	-7.2	-7.2	-6.8	-7.3	-7.5	-9.5	-9.8	-8.0	-7.8	-9.3	-8.9
Auro_4X2R	-9.9	-9.1	-9.8	-8.1	-7.4	-7.1	-7.4	-7.8	-9.2	-9.8	-8.3	-7.8	-9.3	-9.0
Afer_4WD0	-5.7	-5.8	-6.0	-5.7	-6.7	-6.2	-5.8	-7.6	-9.2	-8.8	-7.8	-7.4	-8.3	-8.4
Sent_5AHE	-9.1	-5.4	-6.4	-5.2	-8.0	-7.3	-7.4	-8.8	-10.7	-10.2	-8.7	-8.7	-9.6	-9.9
Ecoli_K12	-9.7	-9.7	-9.2	-6.1	-7.2	-6.6	-6.8	-8.6	-9.5	-9.1	-8.6	-8.2	-9.0	-8.6
Jden_4WUI	-8.3	-8.8	-8.2	-6.8	-6.2	-6.1	-6.0	-6.8	-7.4	-7.6	-7.5	-6.9	-7.6	-7.5
Ctra	-8.2	-8.0	-7.4	-7.2	-6.1	-5.5	-5.4	-6.5	-7.1	-7.1	-7.0	-6.2	-6.8	-6.8
SMg1_trpF	-7.2	-8.8	-8.2	-7.3	-6.2	-5.7	-5.7	-6.9	-6.6	-6.7	-7.3	-6.7	-7.6	-6.9
Aodo_4X2R	-8.5	-8.6	-8.8	-7.3	-7.1	-7.1	-7.1	-7.5	-9.7	-9.4	-7.8	-7.6	-9.6	-8.8

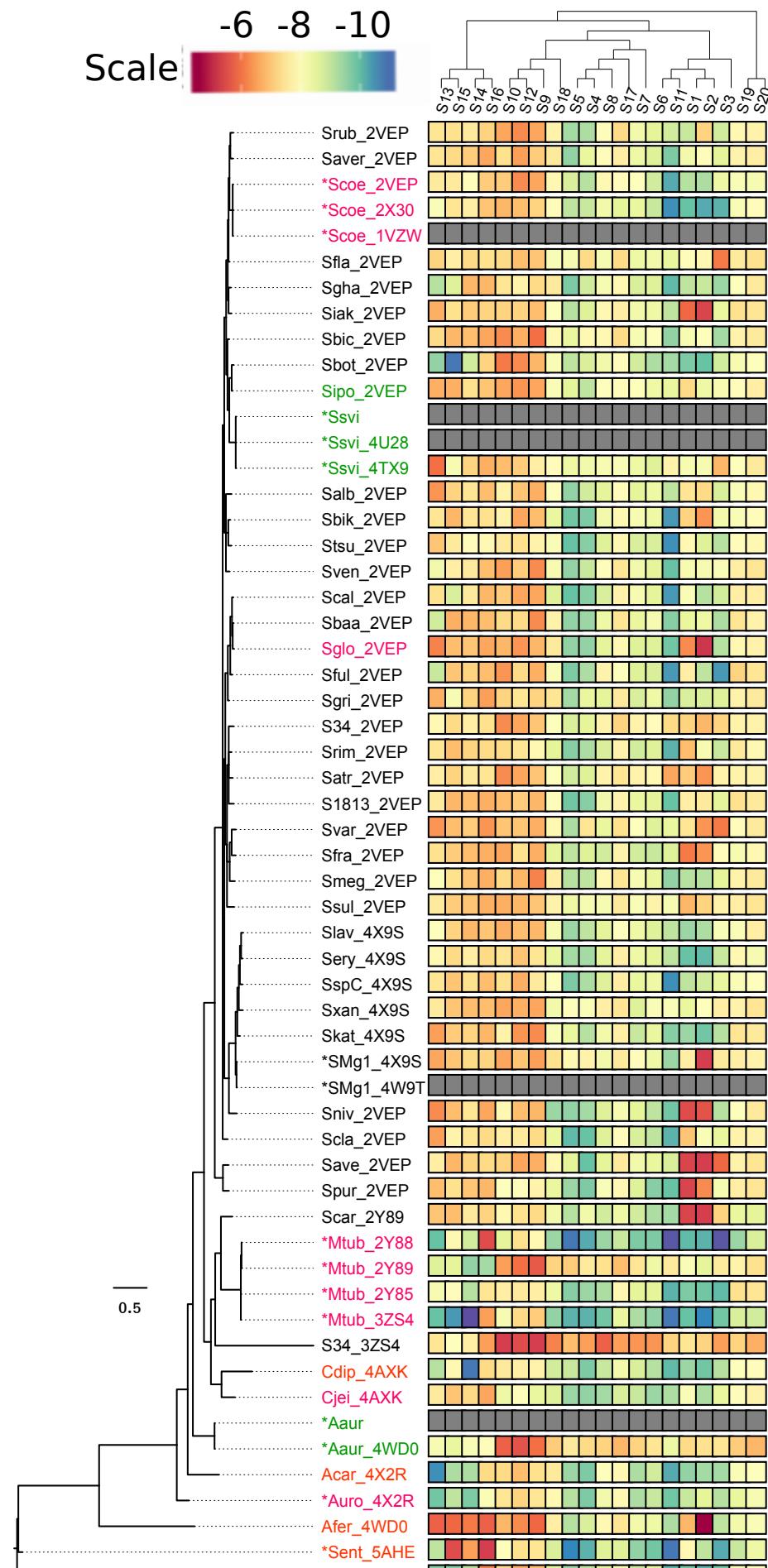
Enzima	S13	S15	S14	S16	S10	S12	S9	S18	S5	S4	S8	S17	S7





Warning: Removed 100 rows containing non-finite values (stat\_boxplot).





organism	Family	$K_M$	$k_{cat}$	$\frac{k_{cat}}{K_M}$	Pre MD	Pos MD	Reference
Afer	HisA	$1.1 \pm 0.2$	$0.05 \pm 0.001$	0.045	-10.1	-12.3	Noda-García
Ecoli	HisA	1.6	4.9	3.1	-9.9	-16	Henn-Sax et
Sent	HisA	$17.0 \pm 0.1$	$7.8 \pm 2.4$	$4.5 \times 10^5$	-10.3	-20.1	Söderholm A
Aaur	PriB	$2.1 \pm 0.5$	$1.8 \pm 0.2$	0.9	-7.4		verduzco-cas
Sipo	PriB	$3.8 \pm 0.2$	$0.82 \pm 0.02$	0.21	-8.2	-14.7	verduzco-cas
SspC	PriB	$11.4 \pm 3.4$	$2.53 \pm 0.74$	0.22	-8.5	-12.7	verduzco-cas
SMg1	PriB	$13.2 \pm 3.4$	$0.92 \pm 0.19$	0.069	-8	-15.2	verduzco-cas
Ssvi	PriB	$3.9 \pm 0.89$	$0.69 \pm 0.04$	0.18	-8.2	-16.7	verduzco-cas
Scoe	PriA	$3.6 \pm 0.7$	$1.3 \pm 0.2$	0.4	-8.4	-15	Noda-García
Sglob	PriA	$4.2 \pm 0.8$	$0.74 \pm 0.03$	0.18	-9.2	-16.7	verduzco-cas
Mtub 2Y85	priA	190.23	$0.012 - 9.7$				Due et al 201
Mtub 3ZS4	priA	?	-9.9				Due et al 201
Auro	priA	$4.0 \pm 0.9$	$0.2 \pm 0.03$	0.04	-9.2		Vazquez-Jua
Cjei	PriA	$2.3 \pm 0.2$	$0.9 \pm 0.08$	0.39	-8.5		Noda-García
Cdip	subHisA	$4.4 \pm 0.5$	$2.6 \pm 0.3$	0.59	-9.2		Noda-García
SMg1 TrpF	TrpF3	-	-	-	-6.9	-9.6	verduzco-cas
Jden	TrpF3	-	-	-	-7.2	-9.4	$16.8 \pm 3.3$
Acar	SubHisA	0.02					Verduzco-Ca
Aodo	SubTrpF	-	-	-			

PriA S7

\$\$

organism	Family	$K_M$	$k_{cat}$	$\frac{k_{cat}}{K_M}$	Pre MD	Pos MD	Reference
Afer	HisA	- -	-	-9.2	-9		Noda-García L et al. (2013)
Ecoli	HisA	- -	-	-9	-11.1		Henn-Sax et al. (2002)
Sent	HisA	- -	-	-9.6	-10.2		Söderholm A et al (2013)
Aaur	PriB	$26.3 \pm 6.3$	$0.37 \pm 0.09$	0.014	-7.1	-	verduzco-castro 2016
Sipo	PriB	$60.8 \pm 1.1$	$8.25 \pm 0.4$	0.14	-8	-8.5	verduzco-castro 2016
SspC	PriB	$149.9 \pm 29$	$1.4 \pm 0.12$	0.009	-8.5	-10.8	verduzco-castro 2016
SMg1	PriB	$129.6 \pm 34$	$0.29 \pm 0.04$	0.0022	-7.5	-11	verduzco-castro 2016
Ssvi	PriB	$24.5 \pm 4.0$	$1.6 \pm 0.29$	0.067	-8	-9.7	verduzco-castro 2016
Scoe	PriA	$5.0 \pm 0.08$	$3.4 \pm 0.09$	0.7	-8	-9.4	Noda-García et al (2013)
Sglob	PriA	$11 \pm 1.0$	$3.8 \pm 0.2$	0.34	-8.7	-9.4	verduzco-castro 2016
Mtub2Y85	priA	21	3.6	0.17	-8.6		Due et al 2011
Mtub3ZS4	priA				-9.3		Due et al 2011 (To be published)
Auro	priA	$23 \pm 6.5$	$0.5 \pm 0.05$	0.02	-9.3		Vazquez-Juarez (2016)
Cjei	PriA	$5.1 \pm 1.0$	$1.6 \pm 0.16$	0.31	-9		Noda-García et al (2013)
Cdip	subHisA	- -	-	-8.8			Noda-García et al (2013)
SMg1 TrpF	TrpF3	$8.4 \pm 1.7$	$10.5 \pm 2.4$	1.25	-7.6	-9	verduzco-castro
Jden	TrpF3	$16.8 \pm 3.3$	$27 \pm 1.6$	1.6	-7.6	-7.7	verduzco-castro
Acar	SubHisA	Na	Na	0.02	Na	Na	Na
Aodo	SubTrpF	- -	-	-	-	Na	Na

\$\$

Con actividad de FolE i.e activa para el compuesto V Adams et al (2014) Genome size vs Total antimash cluster coloured by order

Docker simulation were calculated for Streptomyces enzymes  
Genome size vs Total antimash cluster coloured by order

## 2.5 Latex no Docking

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_de
```

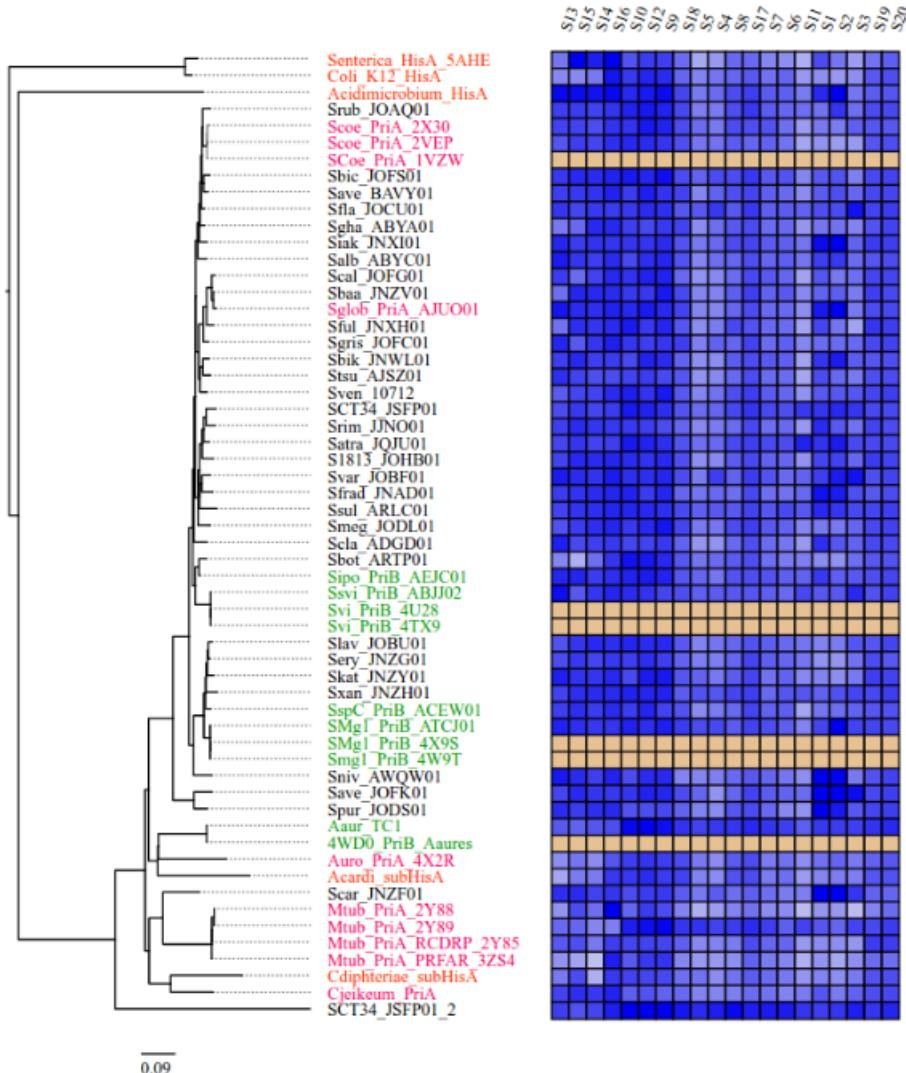


Figure 2.6: Heat Plot PriA Streptomyces vs other substrates



Figure 2.7: Heat Plot TrpF Streptomyces vs other substrates

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()
```

$\text{\TeX}$  is the best way to typeset mathematics. Donald Knuth designed  $\text{\TeX}$  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  $\text{\LaTeX}$  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.

## 2.6 Chemistry 101: Symbols

Chemical formulas will look best if they are not italicized. Get around math mode's automatic italicizing in  $\text{\LaTeX}$  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:  $\text{O}^-$

Subscript:  $\text{CH}_4$

To stack numbers or letters as in  $\text{Fe}_2^{2+}$ , the subscript is defined first, and then the superscript is defined.

Angstrom:  $\text{\AA}$

Bullet:  $\text{CuCl} \bullet 7\text{H}_2\text{O}$

Double Dagger:  $\ddagger$

Delta:  $\Delta$

Reaction Arrows:  $\longrightarrow$  or  $\xrightarrow{\text{solution}}$

Resonance Arrows:  $\leftrightarrow$

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

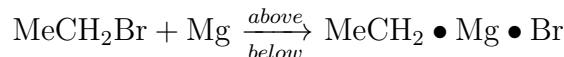
### 2.6.1 Typesetting reactions

You may wish to put your reaction in a figure environment, which means that L<sup>A</sup>T<sub>E</sub>X 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 2.8: Combustion of glucose

### 2.6.2 Other examples of reactions



## 2.7 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<sup>A</sup>T<sub>E</sub>X Symbol Guide (<http://mirror.utexas.edu/ctan/info/symbols/comprehensive/symbols-letter.pdf>).

## 2.8 Biology

You will probably find the resources at <http://www.lecb.ncifcrf.gov/~toms/latex.html> helpful, particularly the links to bst's 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.

```
sessionInfo()
```

```
R version 3.3.2 (2016-10-31)
Platform: x86_64-pc-linux-gnu (64-bit)
Running under: Ubuntu 14.04.5 LTS
```

```
locale:
```

[1] LC_CTYPE=en_US.UTF-8	LC_NUMERIC=C
[3] LC_TIME=es_MX.UTF-8	LC_COLLATE=en_US.UTF-8

```
[5] LC_MONETARY=es_MX.UTF-8      LC_MESSAGES=en_US.UTF-8
[7] LC_PAPER=es_MX.UTF-8        LC_NAME=es_MX.UTF-8
[9] LC_ADDRESS=es_MX.UTF-8       LC_TELEPHONE=es_MX.UTF-8
[11] LC_MEASUREMENT=es_MX.UTF-8   LC_IDENTIFICATION=es_MX.UTF-8

attached base packages:
[1] parallel stats      graphics grDevices utils      datasets methods
[8] base

other attached packages:
[1] xlsx_0.5.7          xlsxjars_0.6.1    rJava_0.9-8
[4] scales_0.4.1         Biobase_2.34.0    BiocGenerics_0.20.0
[7] genstats_0.1.02      RColorBrewer_1.1-2  reshape_0.8.6
[10] plyr_1.8.4           ape_4.0            knitr_1.15.1
[13] ggplot2_2.2.1        dplyr_0.5.0      reedtemplates_0.1
[16] devtools_1.12.0

loaded via a namespace (and not attached):
[1] Rcpp_0.12.9          highr_0.6        tools_3.3.2    digest_0.6.12
[5] evaluate_0.10         memoise_1.0.0    tibble_1.2     gtable_0.2.0
[9] nlme_3.1-131          lattice_0.20-34   DBI_0.5-1     yaml_2.1.14
[13] withr_1.0.2           stringr_1.1.0    rprojroot_1.2  grid_3.3.2
[17] R6_2.2.0              rmarkdown_1.3    magrittr_1.5   backports_1.0.5
[21] htmltools_0.3.5       assertthat_0.1   colorspace_1.3-2 labeling_0.3
[25] stringi_1.1.2         lazyeval_0.2.0   munsell_0.4.3
```

# Chapter 3

## 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[115] . Main differences between this kingdoms are that Archaeal DNA is not arranged in a nucleus as in Eucarya and Archaeal cellular walls are not composed from peptidoglycans as in Bacteria. Archaeal proteins may be highly valuable to biotechnology industry for their great stability due to extreme temperature, PH and salt content conditions on Archeal habitats. Despite no Archaeal Natural products biosynthetic gene clusters (BGC's) has been reported on MiBIG, Archaea do have BGC's, some of them seems to be acquired by horizontal gene transfer (HGT) like methano nrps {search reference}. Other Archeal natural products known are archaeosins, Diketopiperazines, Acyl Homoserine Lactones, Exopolysaccharides, Carotenoids, Biosurfactants, Phenazines and Organic Solutes but this knowledge is not comparable to Bacterial BGC's knowledge[99].

Natural products biosynthetic gene clusters search is actually performed using either *high-confidence/low-novelty or low-confidence/high-novelty* bioinformatic approaches [50]. High confidence methods compares query sequences with previously known BGC's such as nrps or PKS, examples of this algorithms are antiSMASH and clusterfinder [????]. 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 [65]. Also on Actinobacteria antiSMASH analysis on 1245 genomes found 774 different classes of natural products, the same analysis on 876 Archaeal genomes, a full kingdom, identifies only 35 BGC's classes. So either Archaea does not have natural products BGC's or this are not yet known. Next paragraph deals with a possible approach about how natural products BGC's can be find.

Archaea resembled Bacteria in that Archaea uses horizontal gene transfer as a genic interchange mechanism, Archaeal genomes contains operons [118] 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 [118]. 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 entirelyyy based on previous knowledge of BGC's sequences if evolutionary logic behave on Archaea as on bacteria, new BGC's classes may 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

### **3.1 Tables**

Table 3.1: 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.

### 3.1.1 Expansions BoxPlot by metabolic family

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

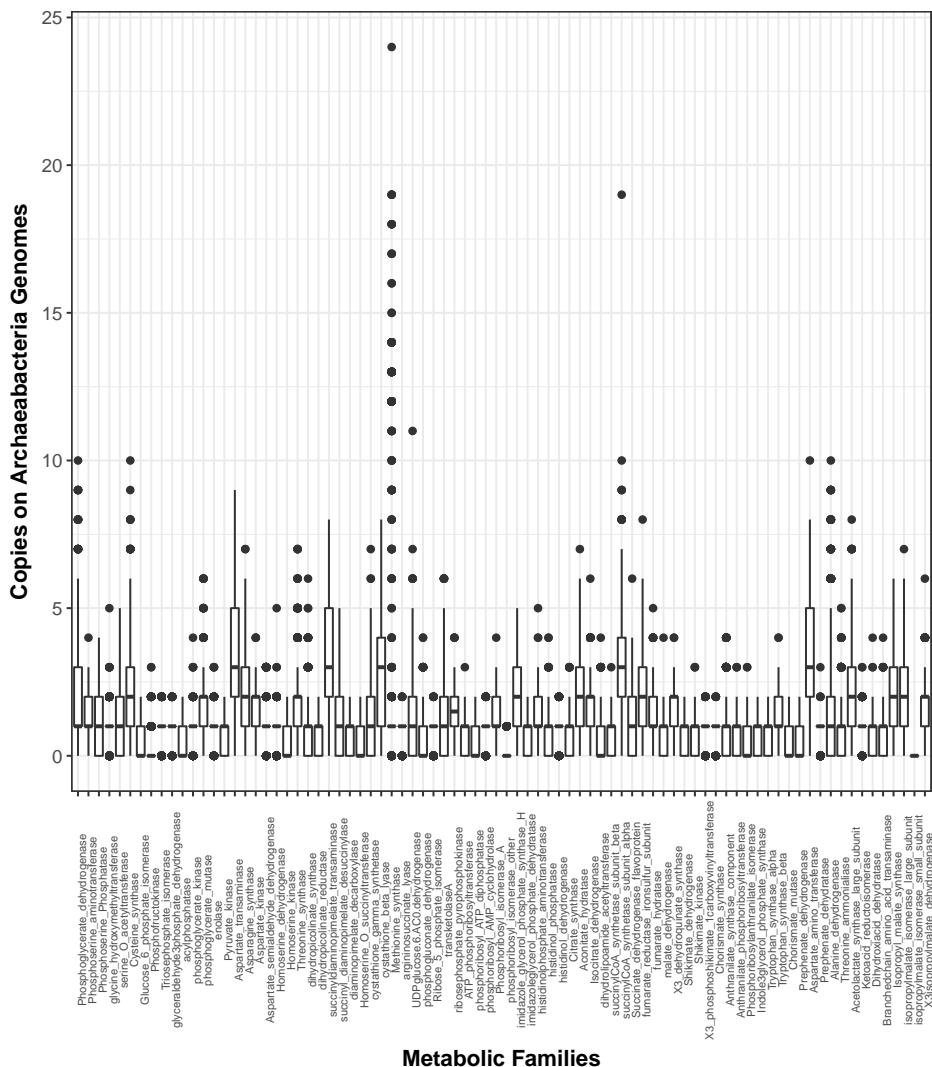


Figure 3.1: Expansions Boxplot

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

### 3.1.2 Expansions BoxPlot by metabolic family by phylum

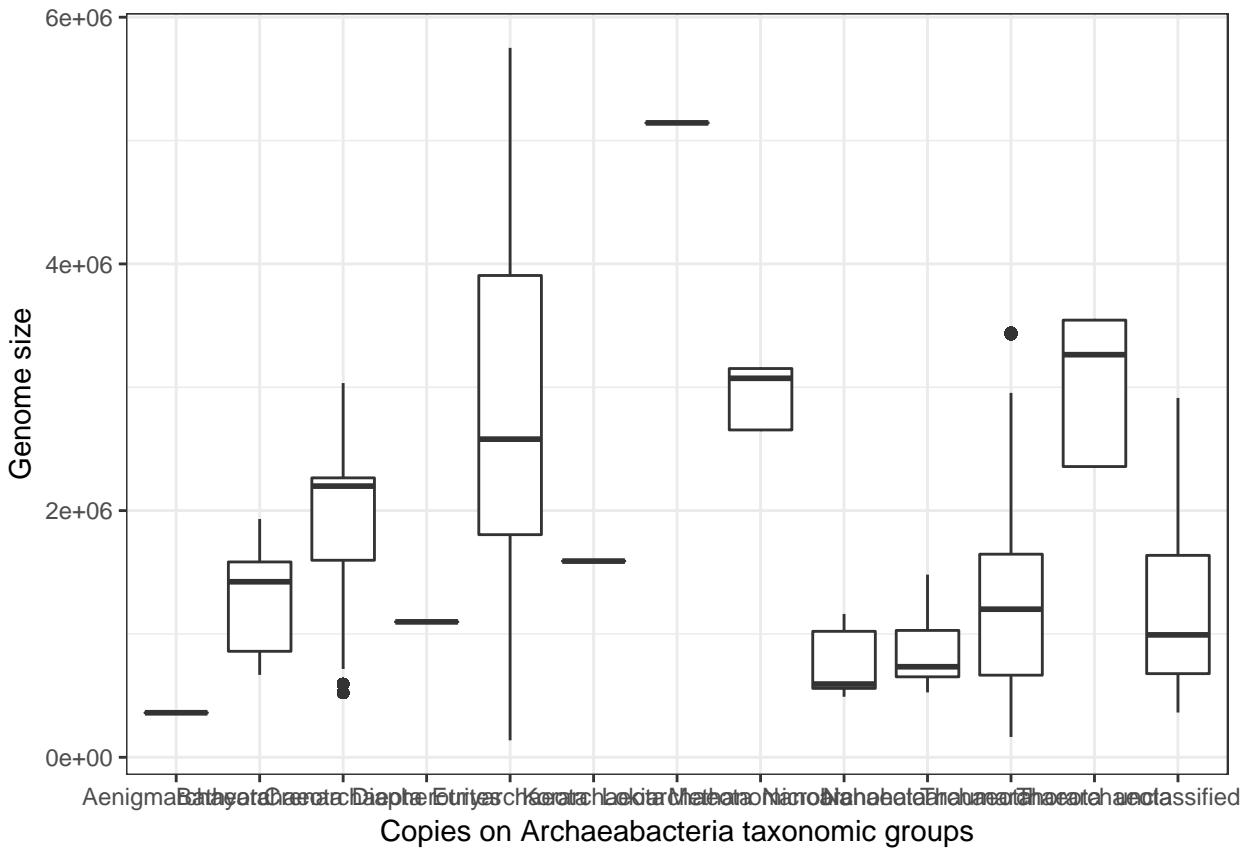
```
#+ geom_jitter()
#aes(fill = factor(vs))

ArchaeasTotalBP.m<-merge(ArchaeasHeatPlot,ArchaeasTaxa,by.x="RastId",by.y="RastId") ## w
ArchaeasHeatPlotBP.m <- melt(ArchaeasTotalBP.m,id =c("RastId","Name","SuperPhylum","Phyl
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

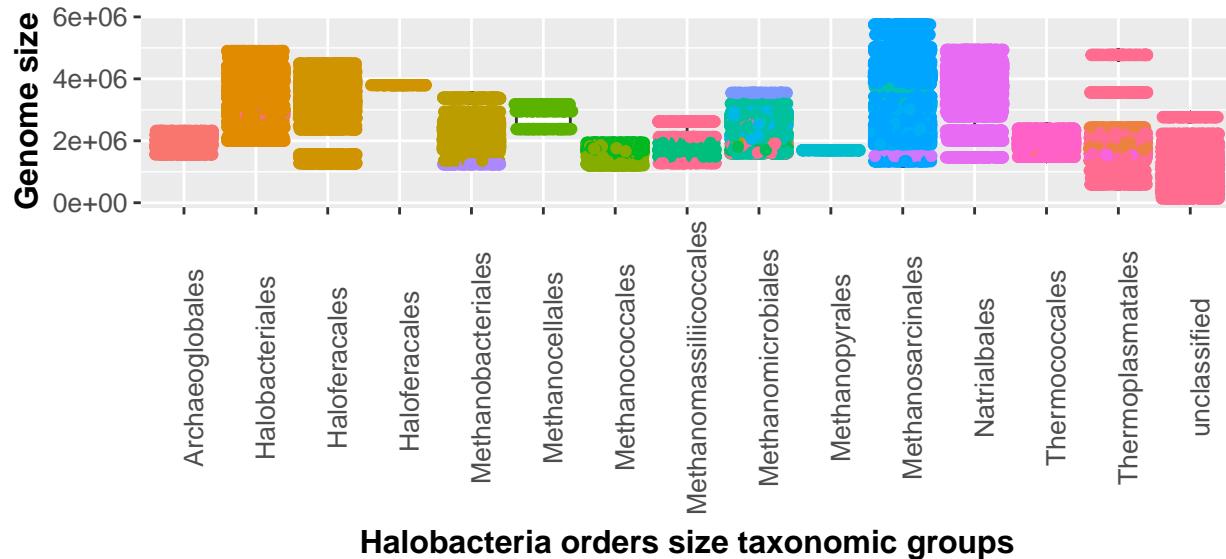
#3PGA_AMINOACIDS
#Glycolysis
#OXALACETATE_AMINOACIDS
#R5P_AMINOACIDS
#TCA
#E4P_AMINO_ACIDS
#PYR_THR_AA

## Genome size
ggplot(ArchaeasHeatPlotBP.m, aes(x=ArchaeasHeatPlotBP.m$Phylum, y=ArchaeasHeatPlotBP.m$S
```



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

## Halobacteria
MetFam_BP.m=subset(ArchaeaHeatPlotBP.m, Phylum=="Euryarchaeota")
ggplot(MetFam_BP.m, aes(x=MetFam_BP.m$Order, y=MetFam_BP.m$Size))+ geom_boxplot()
```

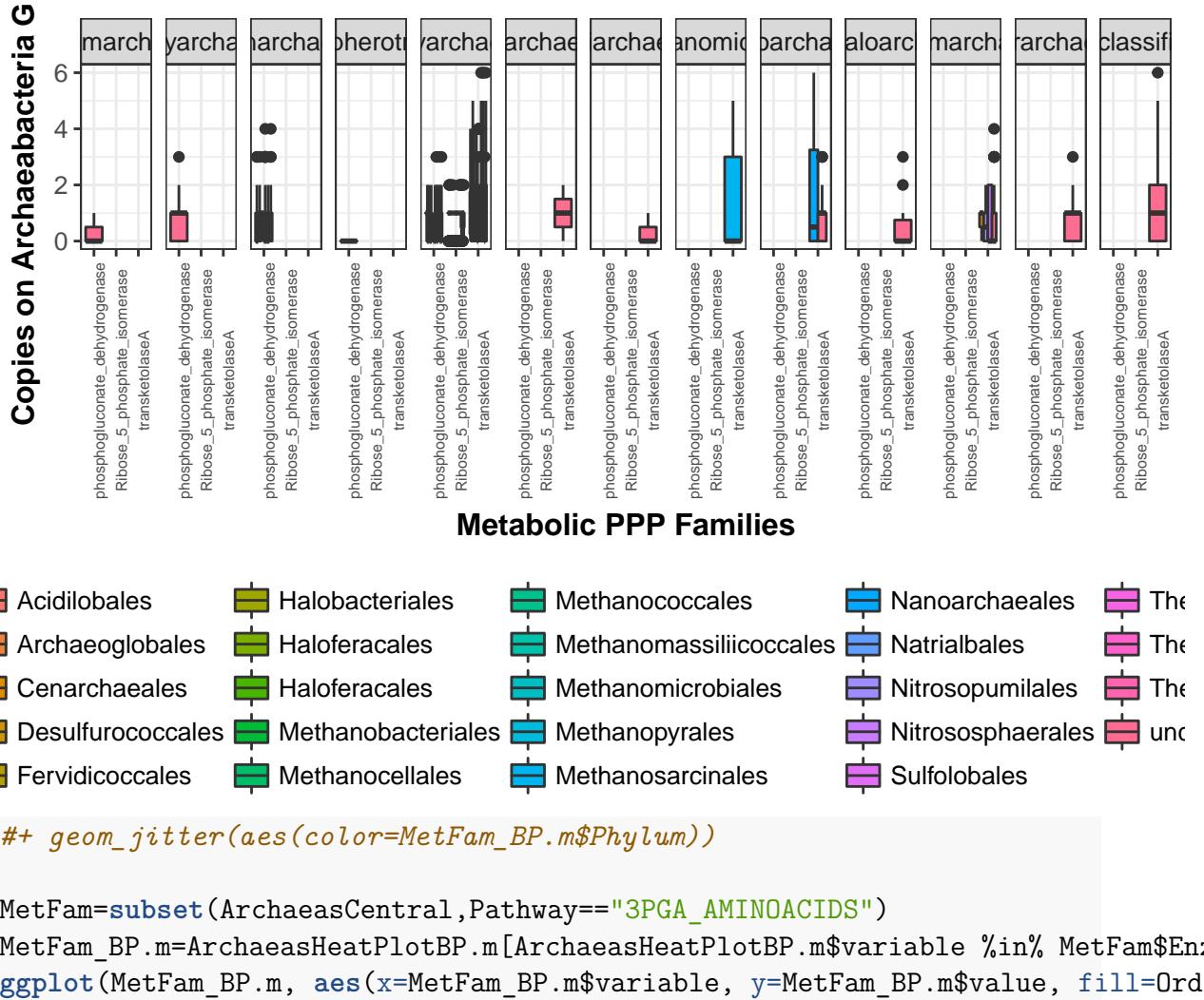


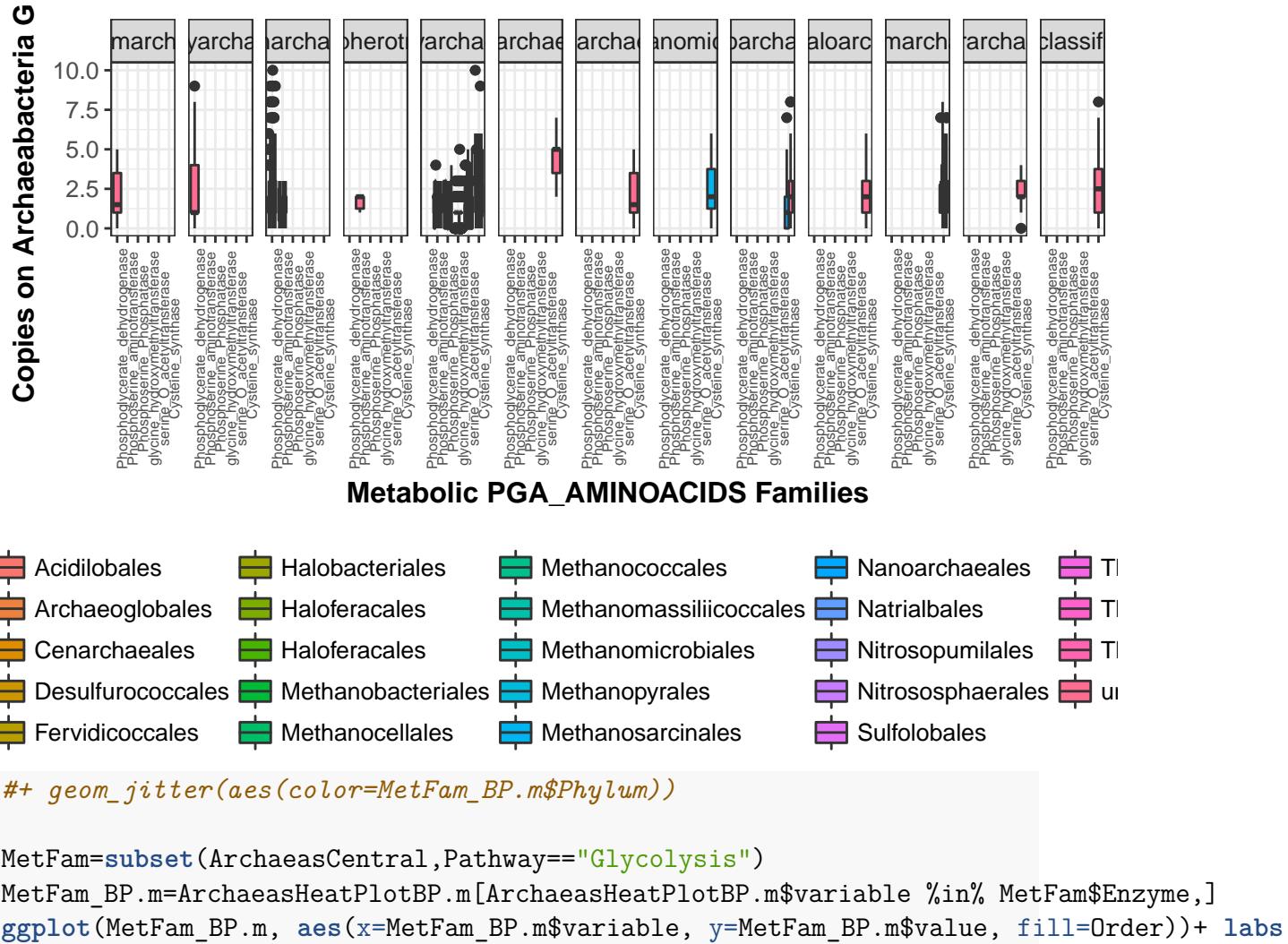
haeoglobaceae	● Methanocalculaceae	● Methanomassiliicoccaceae	● Methanosaetaceae
roplasmaceae	● Methanocaldococcaceae	● Methanomicrobiaceae	● Methanosarcinaceae
obacteriaceae	● Methanocellaceae	● Methanoperedenaceae	● Methanospirillaceae
oferacaceae	● Methanococcaceae	● Methanopyraceae	● Methanothermaceae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

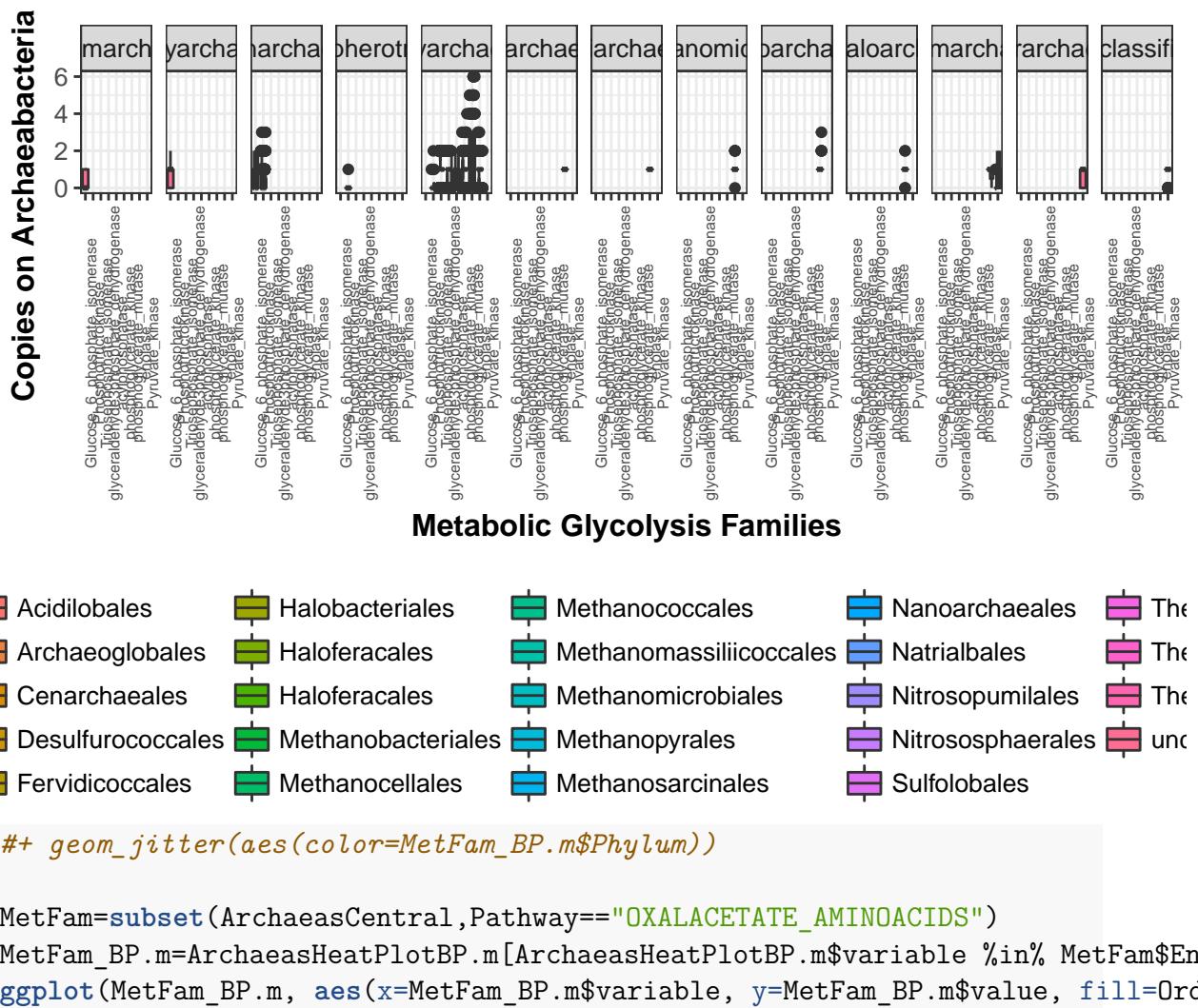
```
#MetFam_BP.m=subset(ArchaeaHeatPlotBP.m,Family=="Methanosarcinaceae")
#ggplot(MetFam_BP.m, aes(x=MetFam_BP.m$Size, y=MetFam_BP.m$value))
#+theme(plot.title = element_text(size = 14, face = "bold"), text = element_text(size = 12))

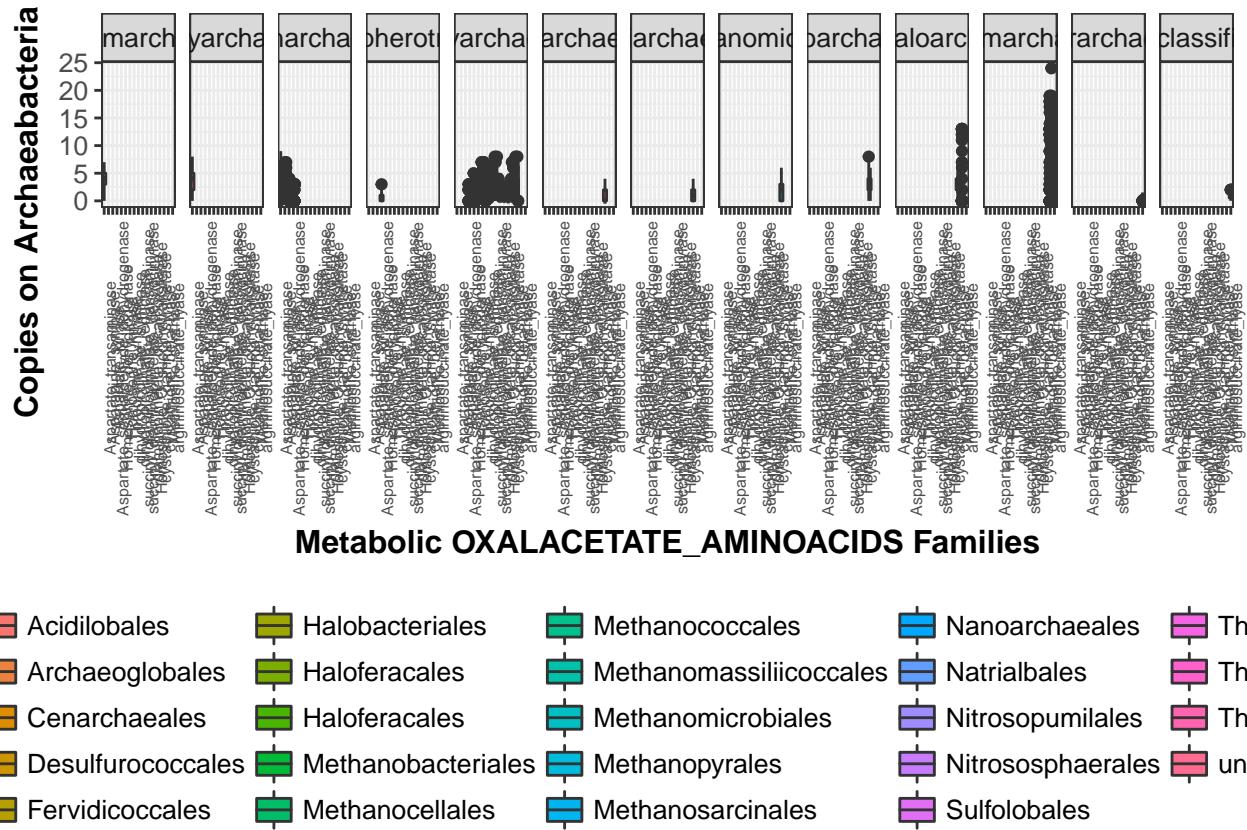
#geom_jitter(aes(color=ArchaeaHeatPlotBP.m$Phylum))# + facet_grid(. ~ Phylum)+theme()

## Metabolic Pathways
MetFam=subset(ArchaeaCentral,Pathway=="PPP")
MetFam_BP.m=ArchaeaHeatPlotBP.m[ArchaeaHeatPlotBP.m$variable %in% MetFam$Enzyme,]
ggplot(MetFam_BP.m, aes(x=MetFam_BP.m$variable, y=MetFam_BP.m$value, fill=Order))+ labs(x="Metabolic Pathways", y="Number of Enzymes")
```



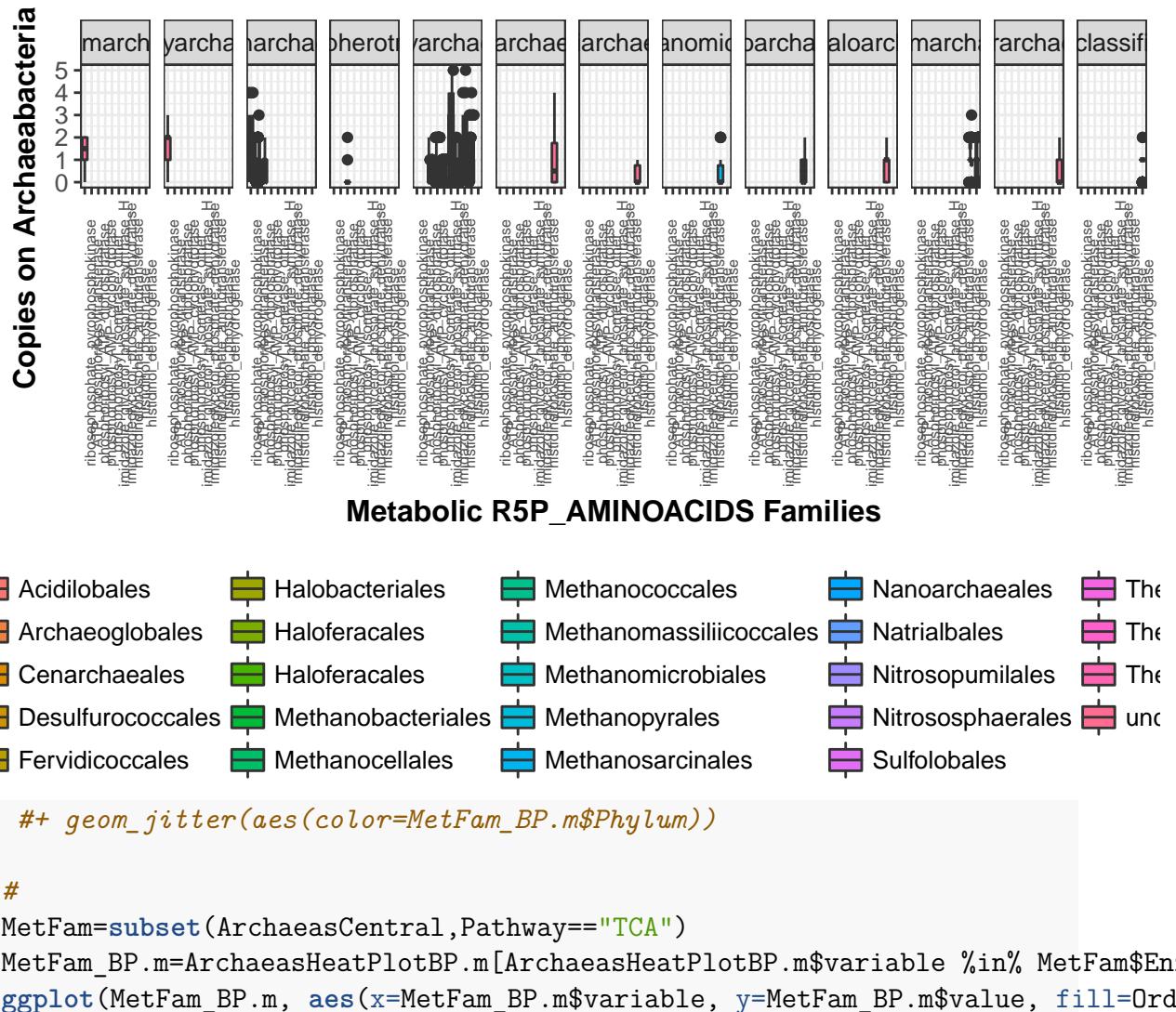


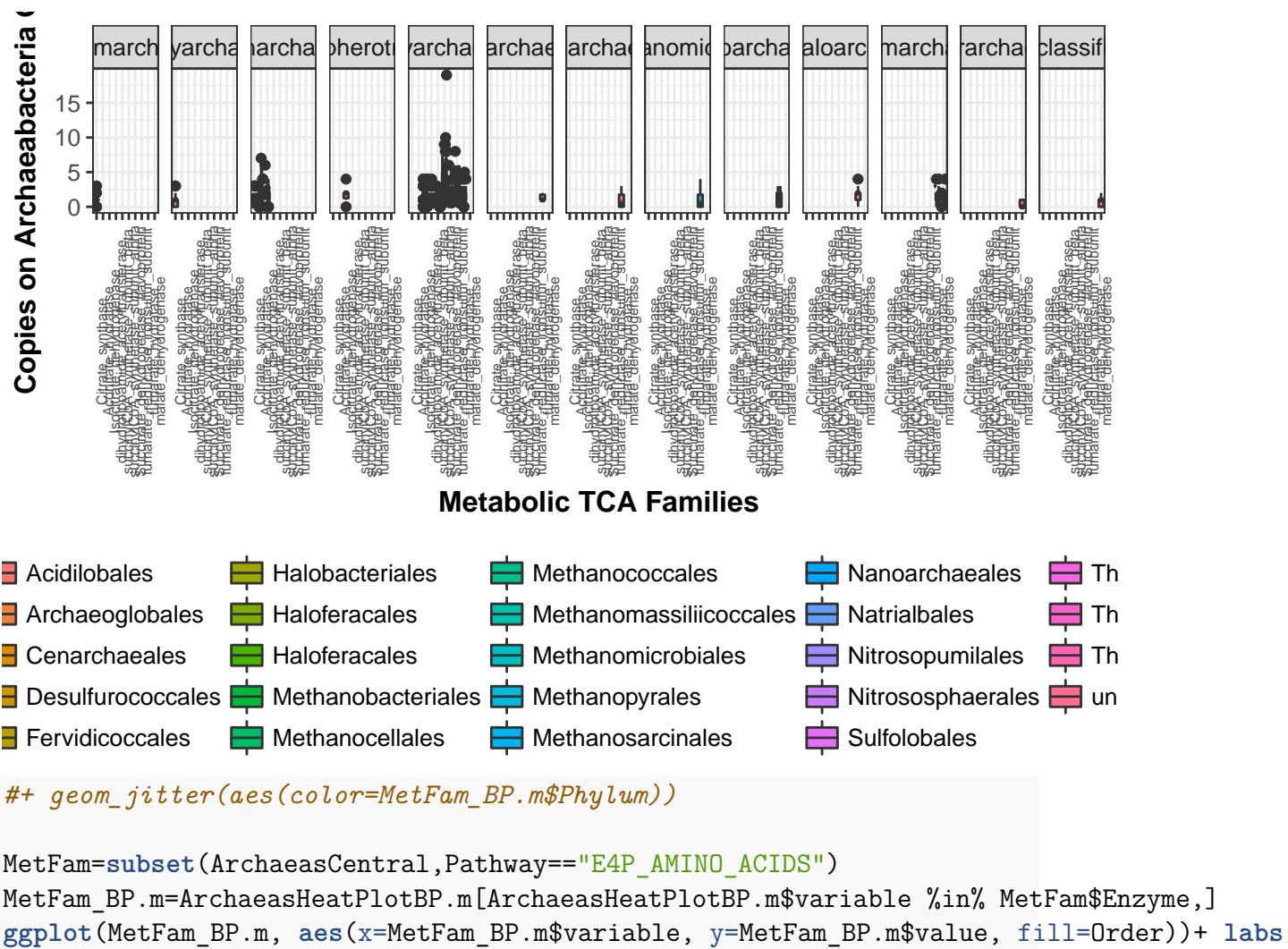


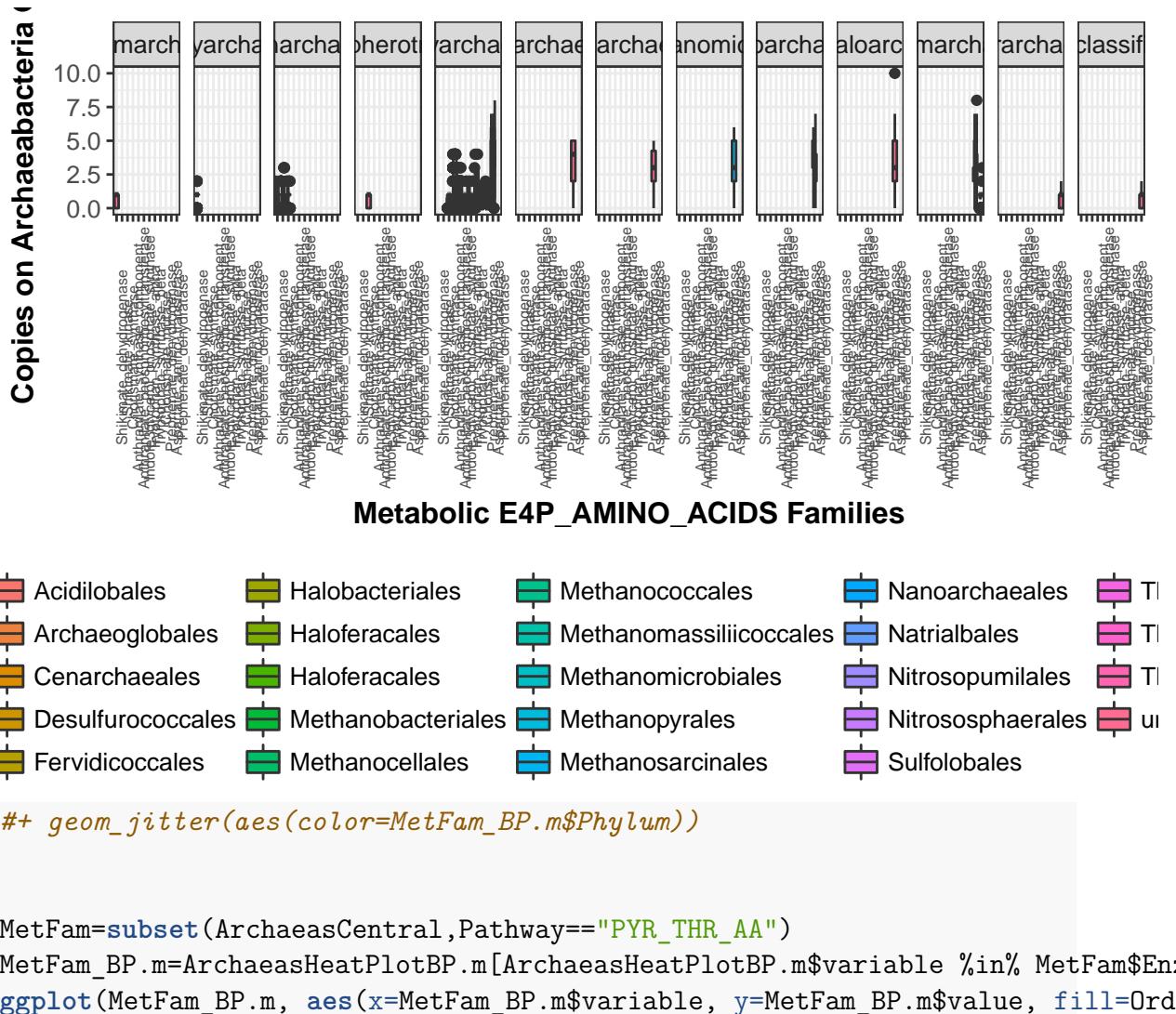


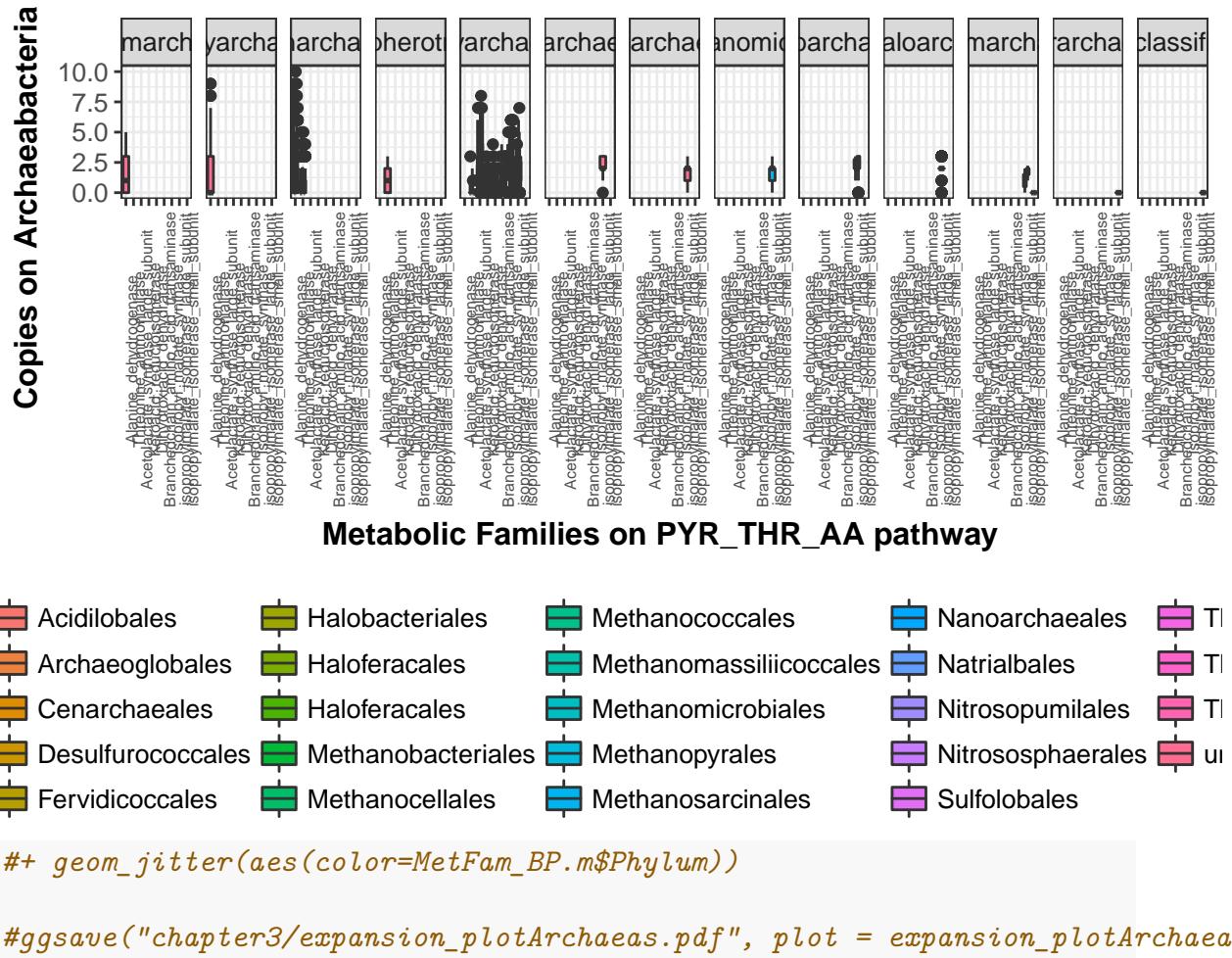
```
#+ geom_jitter(aes(color=MetFam_BP.m$Phylum))

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









## 3.2 Central pathway expansions

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

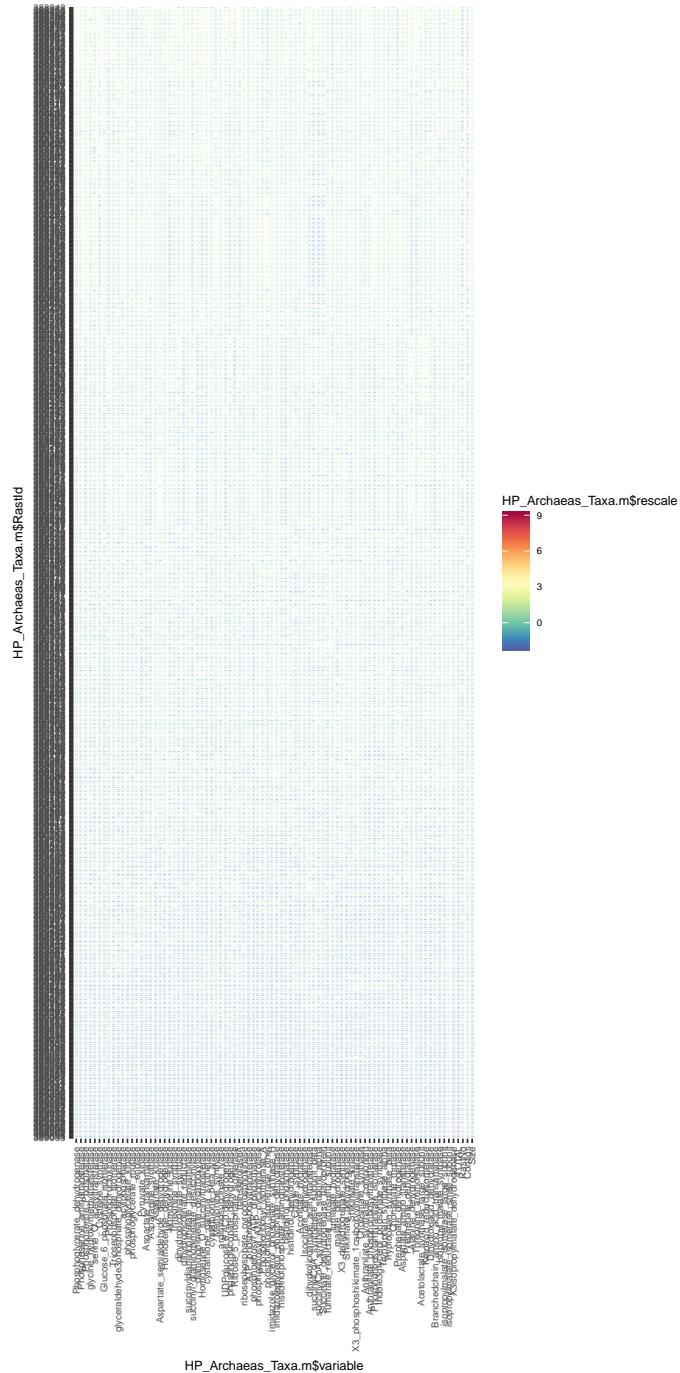


Figure 3.2: Archaeas Heatplot

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

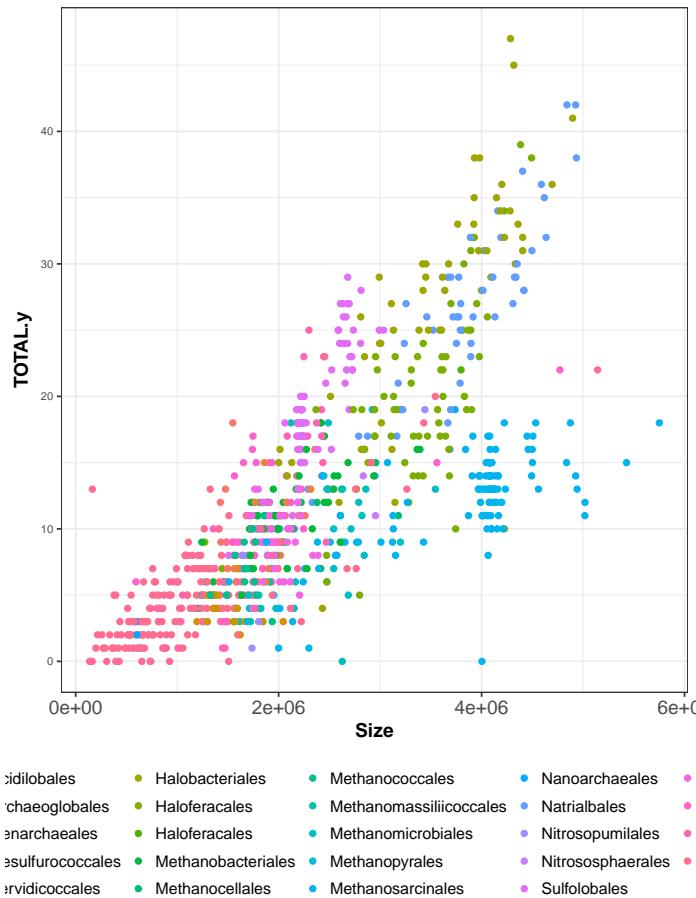


Figure 3.3: Correlation between Archaea genome size and antismash Natural products detection colored by Order

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

Genome size vs Total antismash cluster detected splitted by order

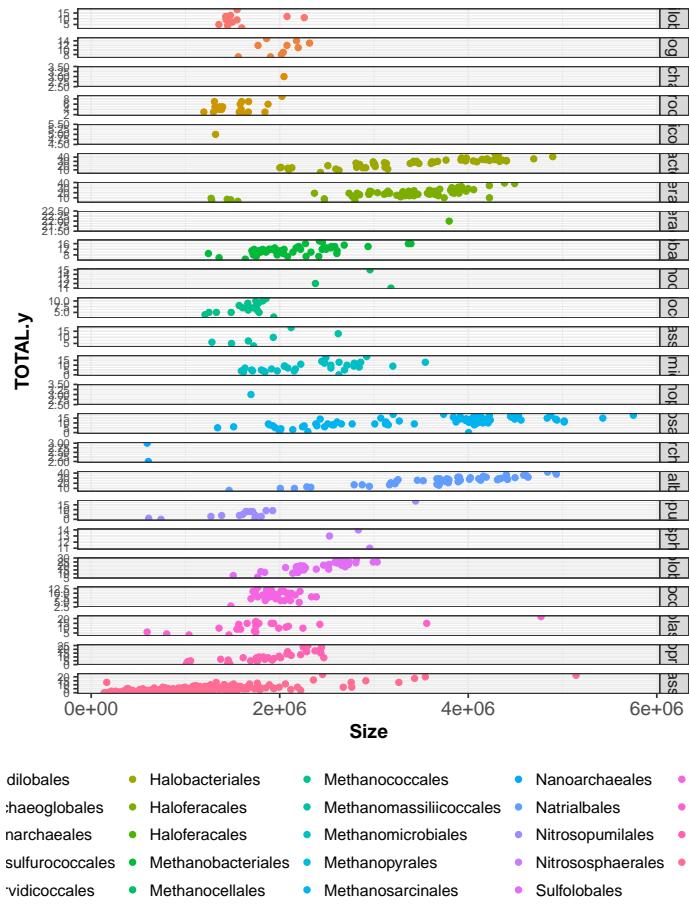


Figure 3.4: Correlation between Archaea genome size and antismash Natural products detection grided by Order

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

### 3.3.2 Correlation between genome size and Central pathway expansions

Genome size vs Total central pathway expansion coloured by order

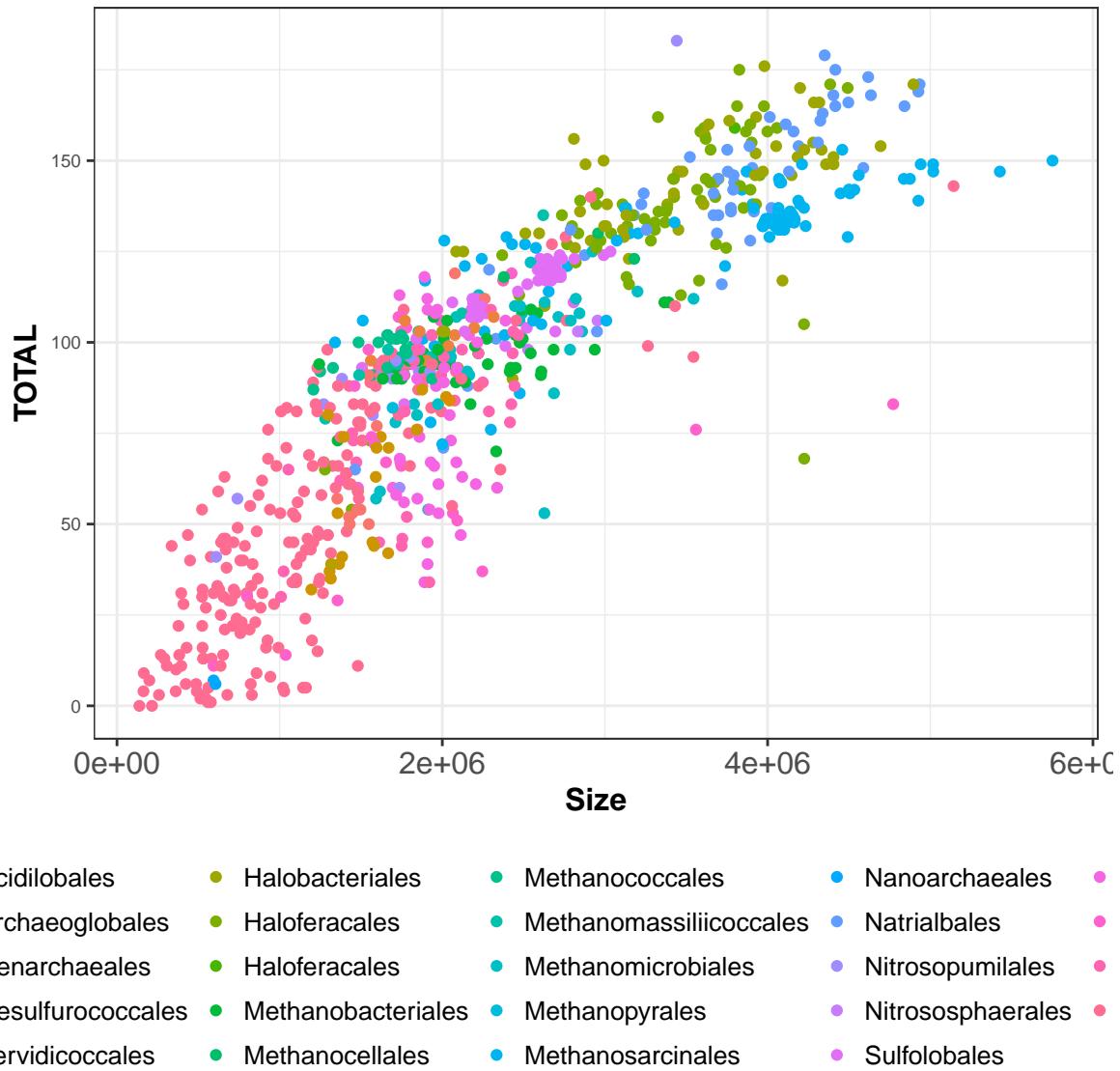


Figure 3.5: Correlation between Archaea genome size and central pathway expansions

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

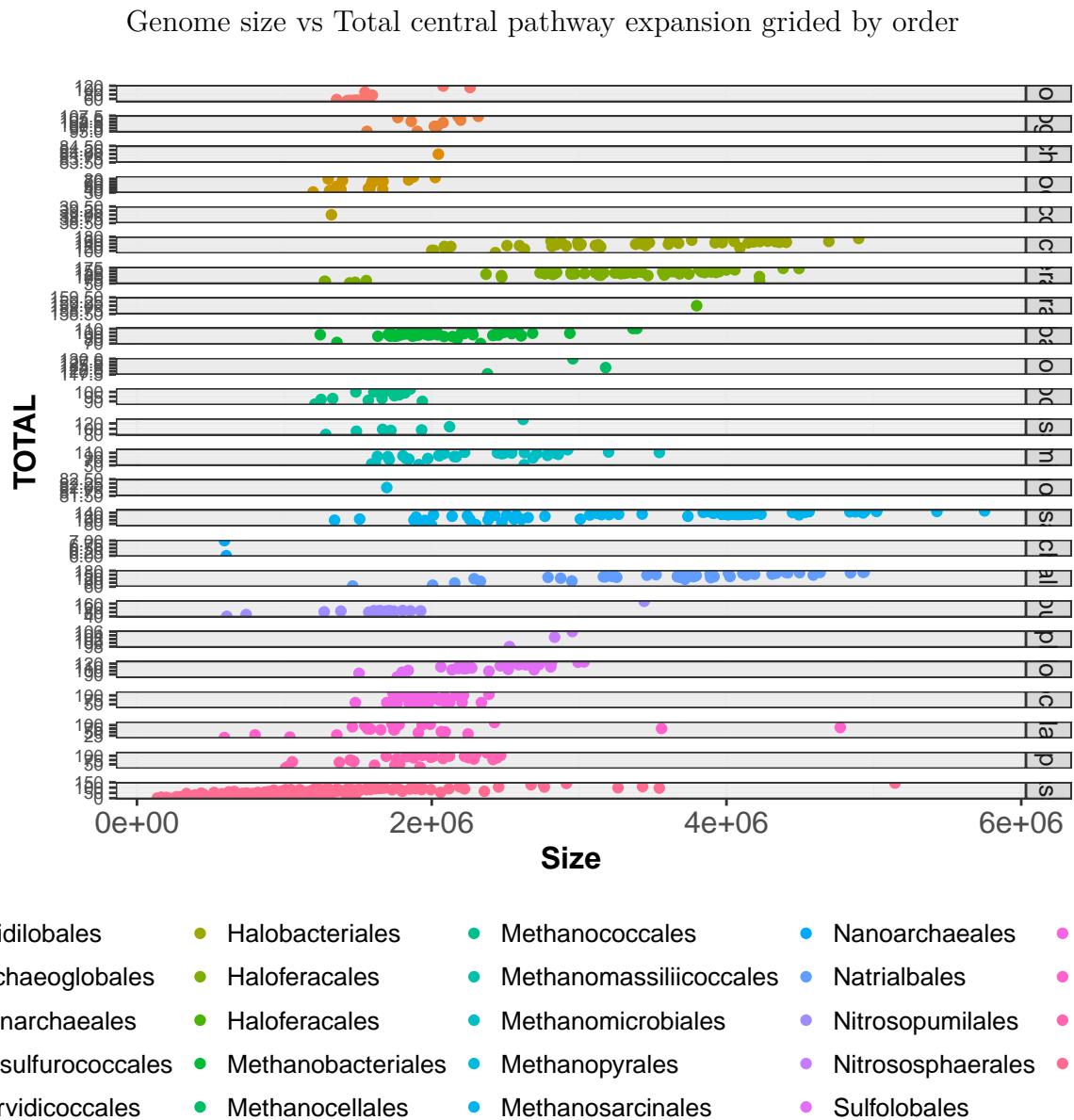


Figure 3.6: Correlation between Archaea genome size and central pathway expansions grided by order

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

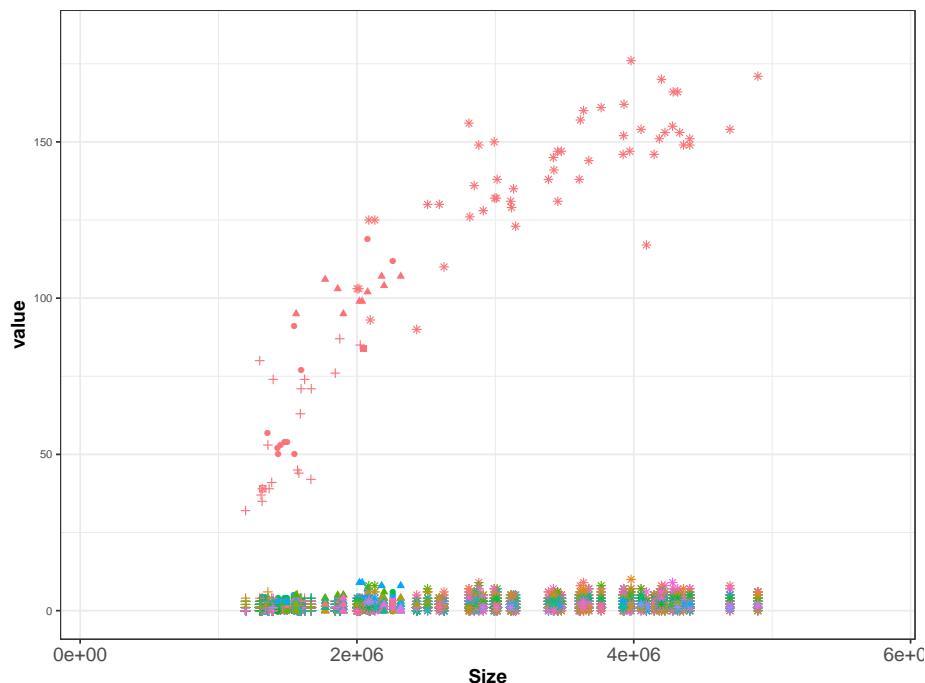


Figure 3.7: Correlation between Archaeas Genome size vs Total central pathway expansion coloured by metabolic Family

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

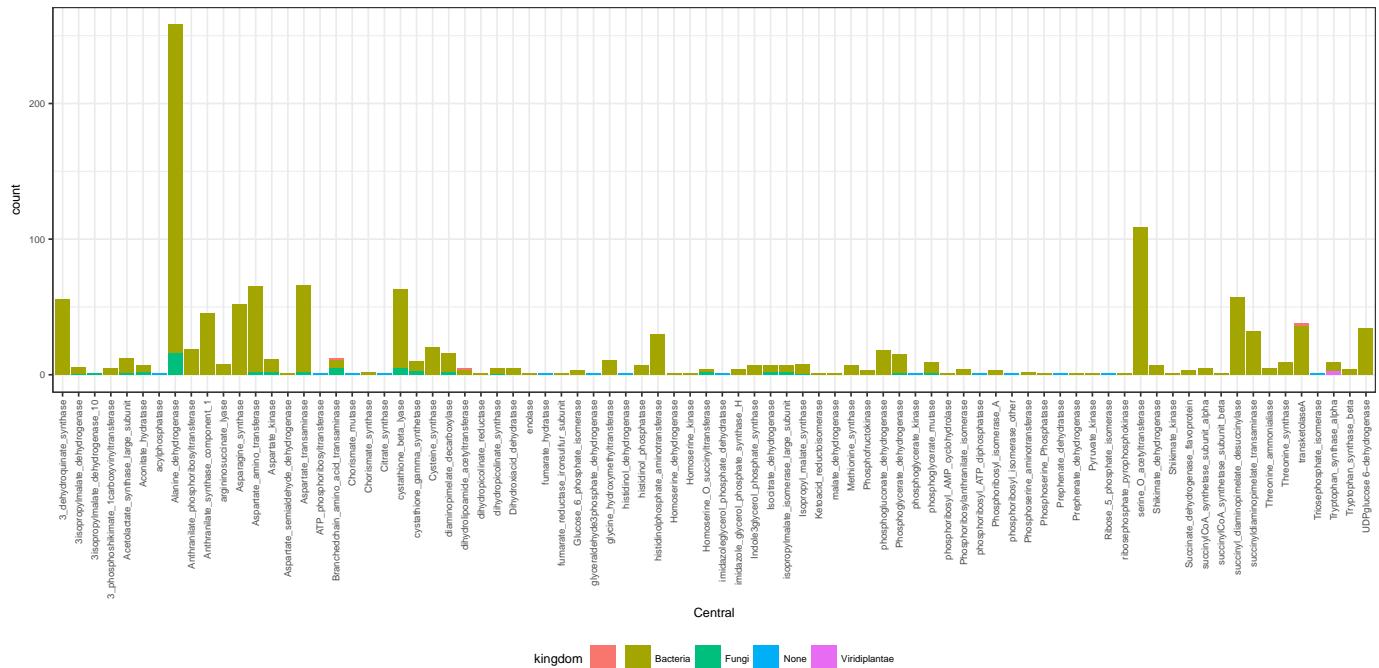


Figure 3.8: Archaeas Recruitmens on central families coloured by kingdom

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

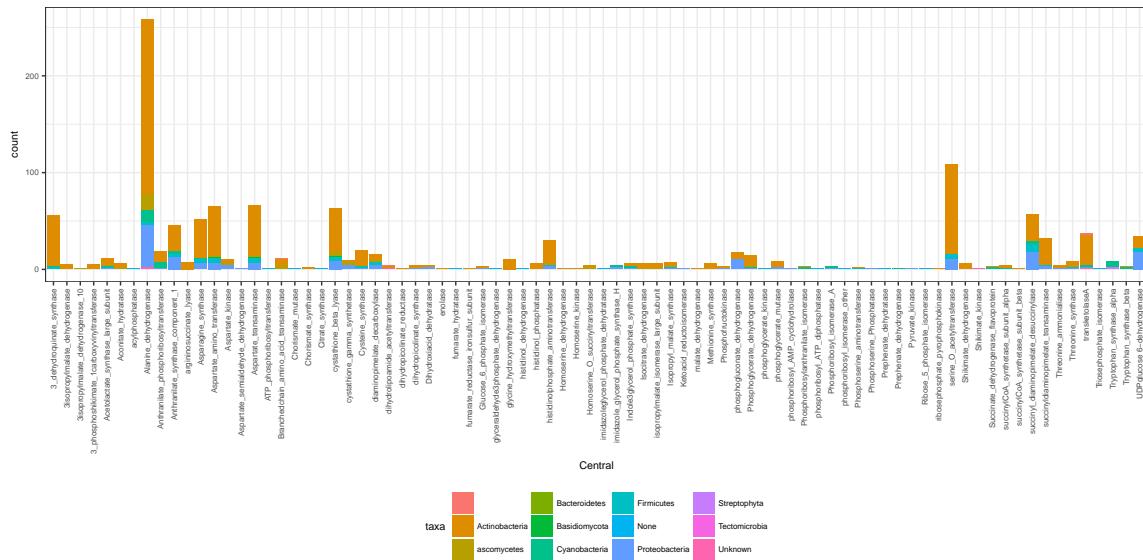


Figure 3.9: Archaeas Recruitmens on central families coloured by taxonomy

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

### 3.5 Archaeas AntiSMASH

Taxonomical diversity on Archaeasbacteria Data

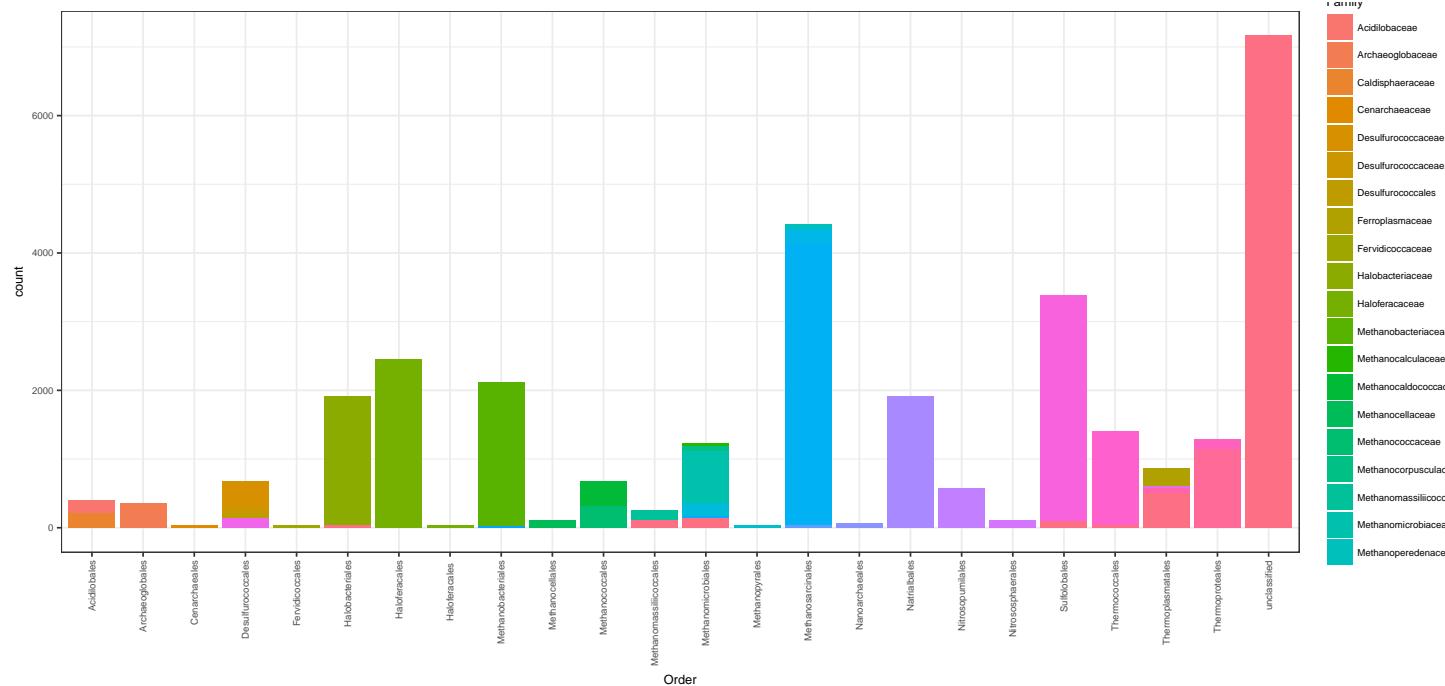


Figure 3.10: Archaea Diversity

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

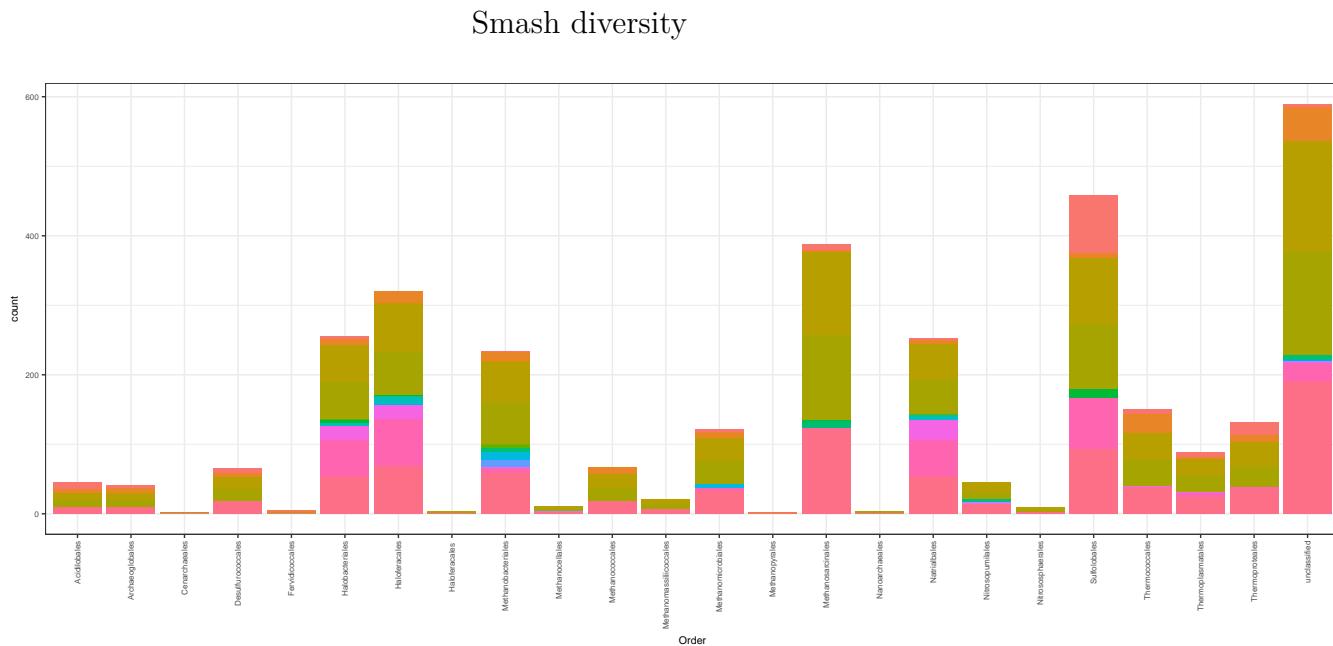


Figure 3.11: Archaeas Smash Taxonomical Diversity

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 antismash cluster detected  
coloured by order

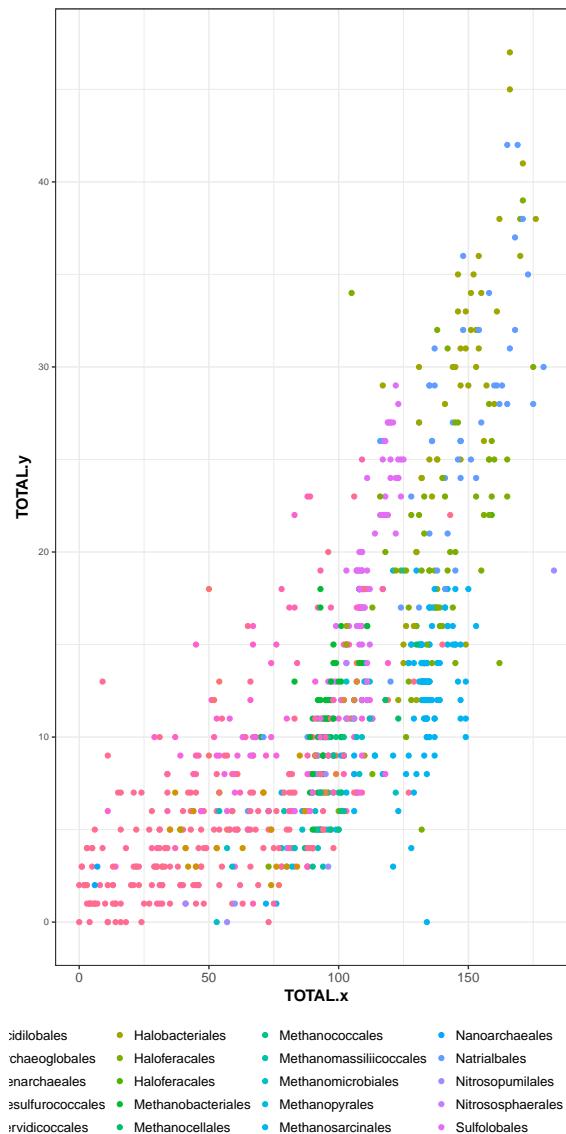


Figure 3.12: Correlation between Archaea's central pathway expansions and antismash Natural products detection

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

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

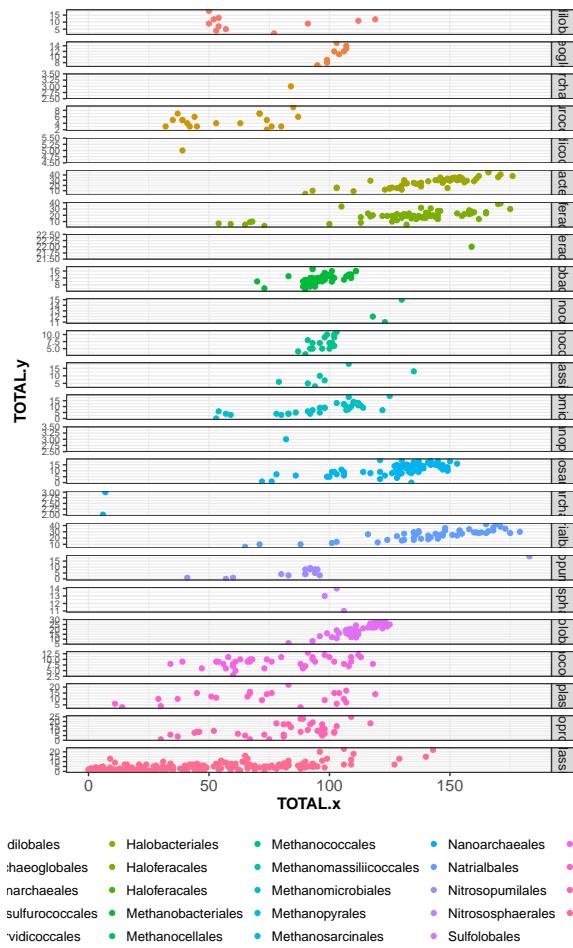


Figure 3.13: Correlation between Archaeas central pathway expnsions and antismash NP's clusters splitted by order plot

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

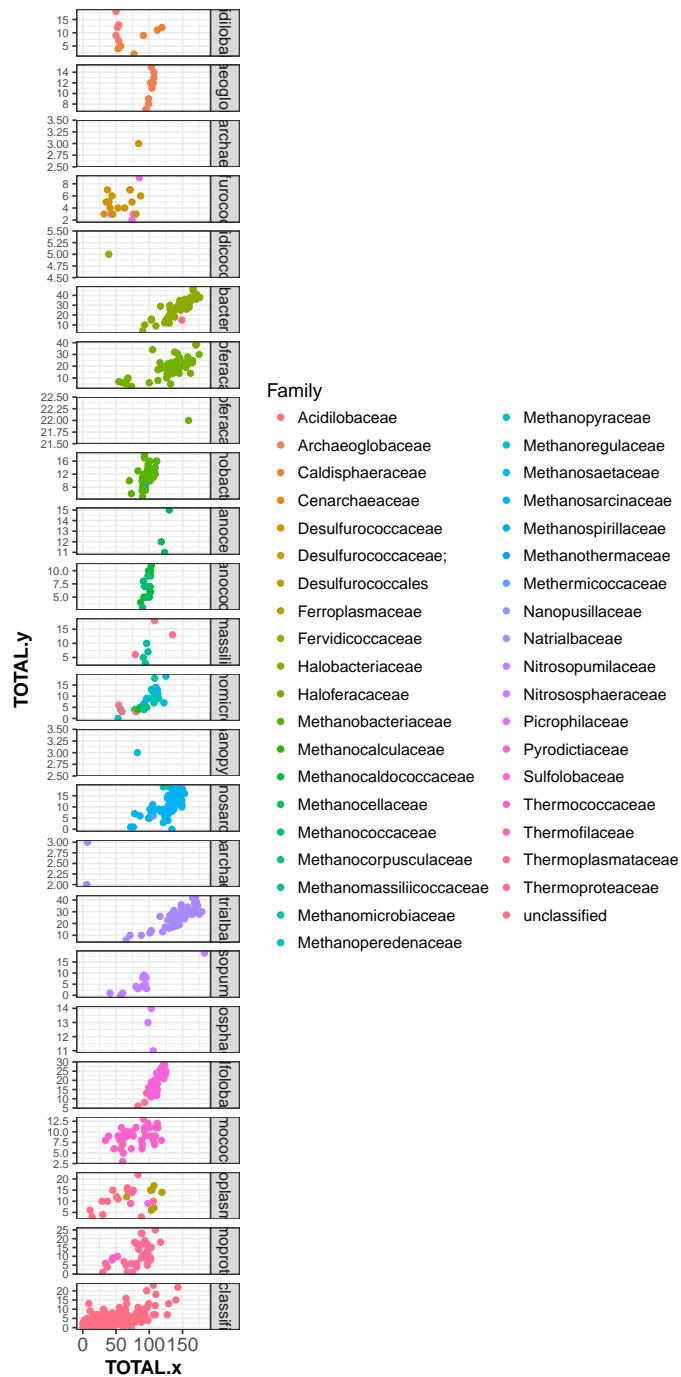


Figure 3.14: Archaeas Natural products by family

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

### 3.6 Selected trees from EvoMining

Phosphoribosyl\_isomerase\_3 family  
Figure from EvoMining

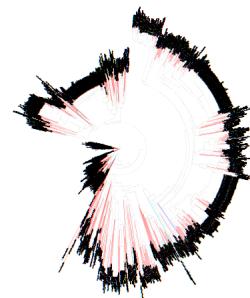


Figure 3.15: Phosphoribosyl isomerase A EvoMiningtree

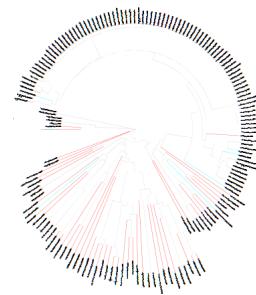


Figure 3.16: Phosphoribosyl isomerase other EvoMiningtree

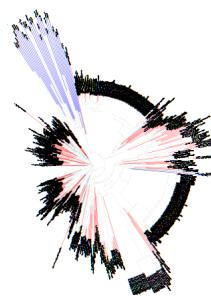


Figure 3.17: Phosphoribosyl anthranilate isomerase EvoMiningtree

## 3.7

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* [117] ## Footnotes and Endnotes

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

*R Markdown* uses *pandoc* (<http://pandoc.org/>) to build its bibliographies. One nice caveat of this is that you won't have to do a second compile to load in references as standard L<sup>A</sup>T<sub>E</sub>X requires. To cite references in your thesis (after creating your bibliography database), place the reference name inside square brackets and precede it by the “at” symbol. For example, here's a reference to a book about worrying: [140]. 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>)<sup>2</sup>. There are three pages on this topic: *bibtex* (which talks about using BibTeX, at <http://web.reed.edu/cis/help/latex/bibtex.html>), *bibtexstyles* (about how to find and use the bibliography style that best suits your needs, at <http://web.reed.edu/cis/help/latex/bibtexstyles.html>) and *bibman* (which covers how to make and maintain a bibliography by hand, without BibTeX, at <http://web.reed.edu/cis/help/latex/bibman.html>). The last page will not be useful unless you have only a few sources.

If you look at the YAML header at the top of the main .Rmd file you can see that we can specify the style of the bibliography by referencing the appropriate csl file. You

---

<sup>1</sup>footnote text

<sup>2</sup>[141]

can download a variety of different style files at <https://www.zotero.org/styles>.

Make sure to download the file into the csl folder.

### Tips for Bibliographies

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

## 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](mailto:data@reed.edu)) with your suggestions. We love to see people using *R Markdown* for their theses, and are happy to help.

---

<sup>3</sup>[142]



# 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.

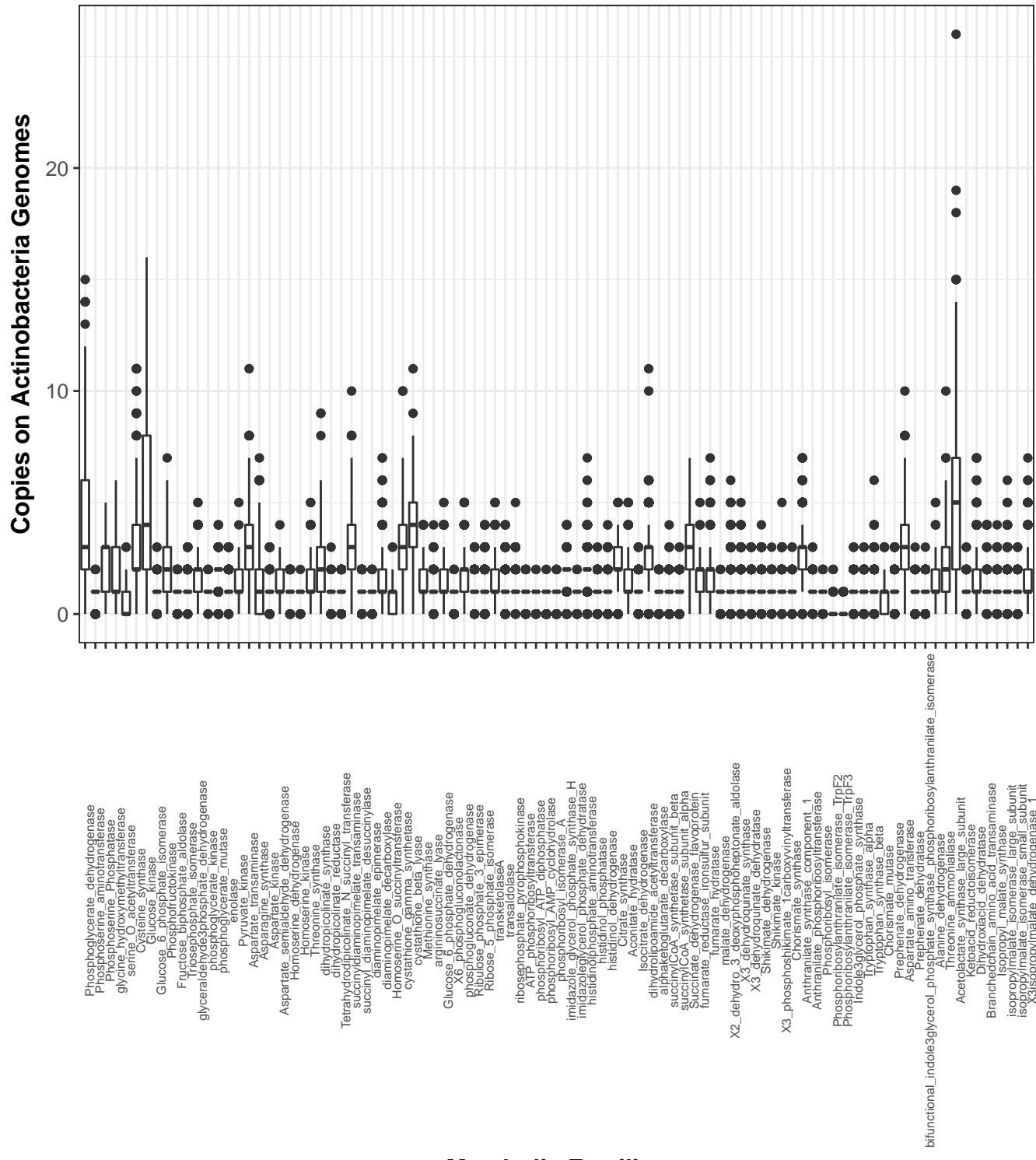
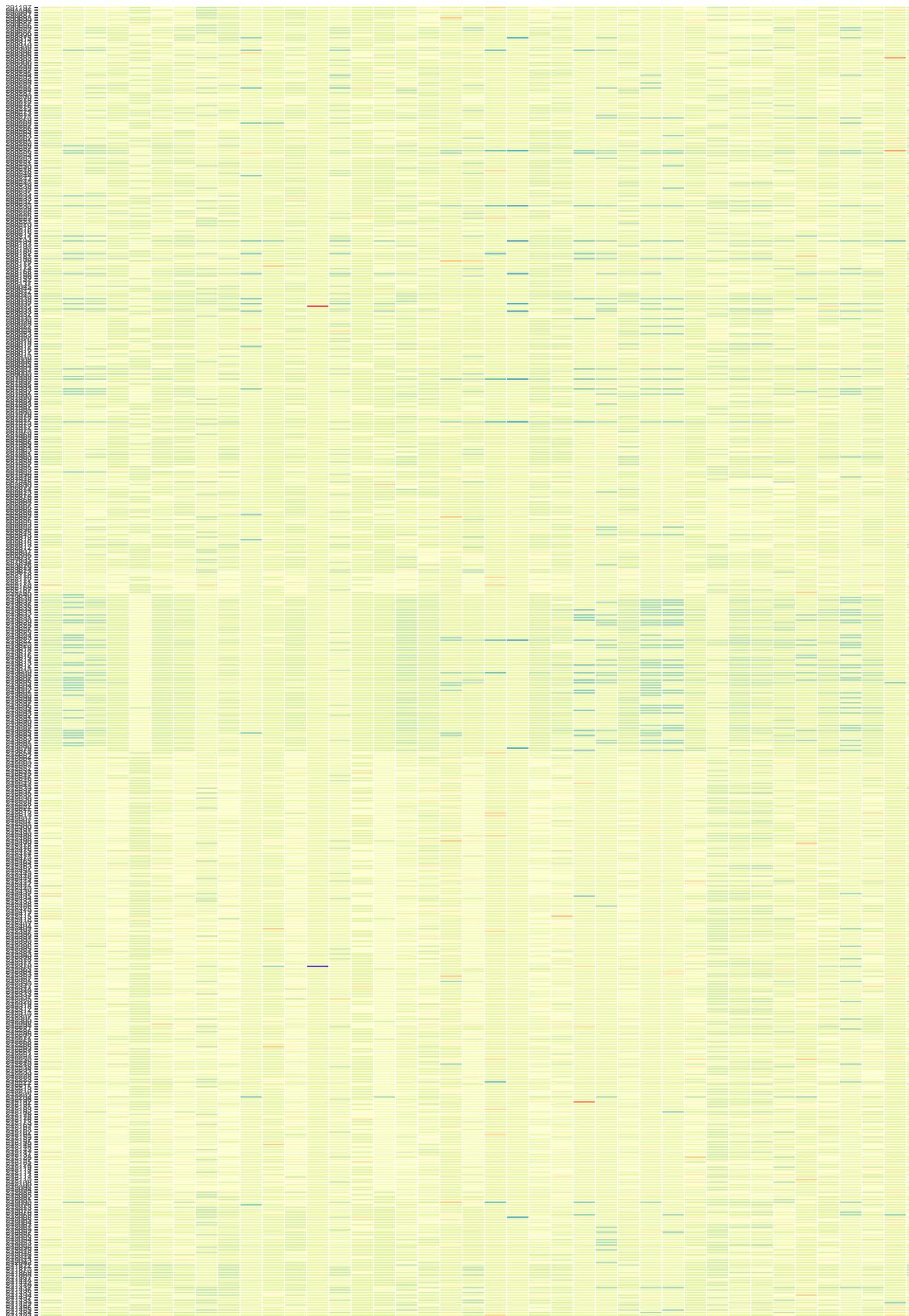


Figure 4.1: Expansions Boxplot

## 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		
287981		
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

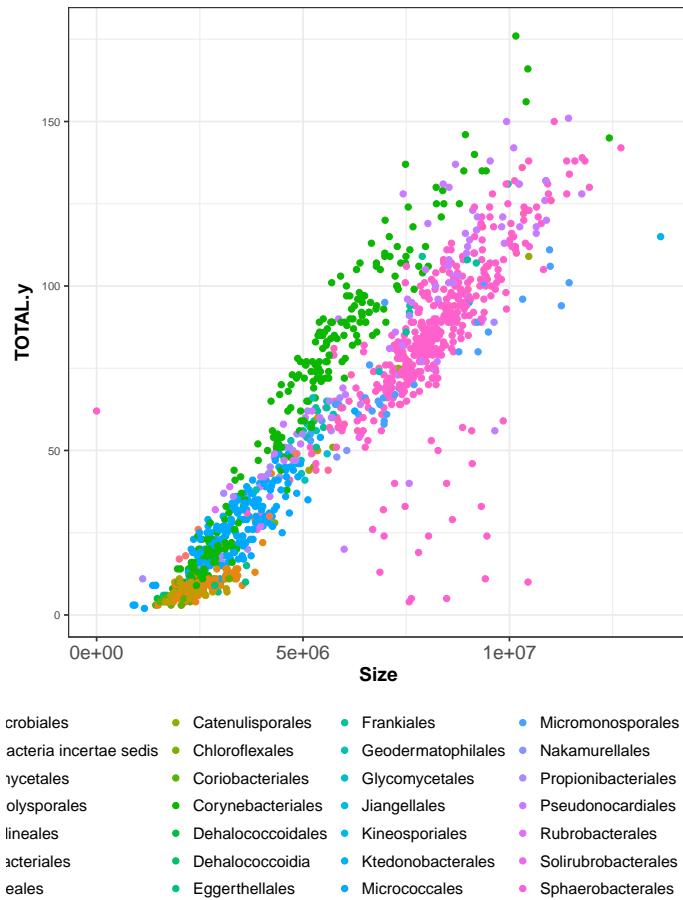


Figure 4.4: Correlation between Actinos genome size and antismash Natural products detection colored by Order

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

Genome size vs Total antismash cluster detected splitted by order

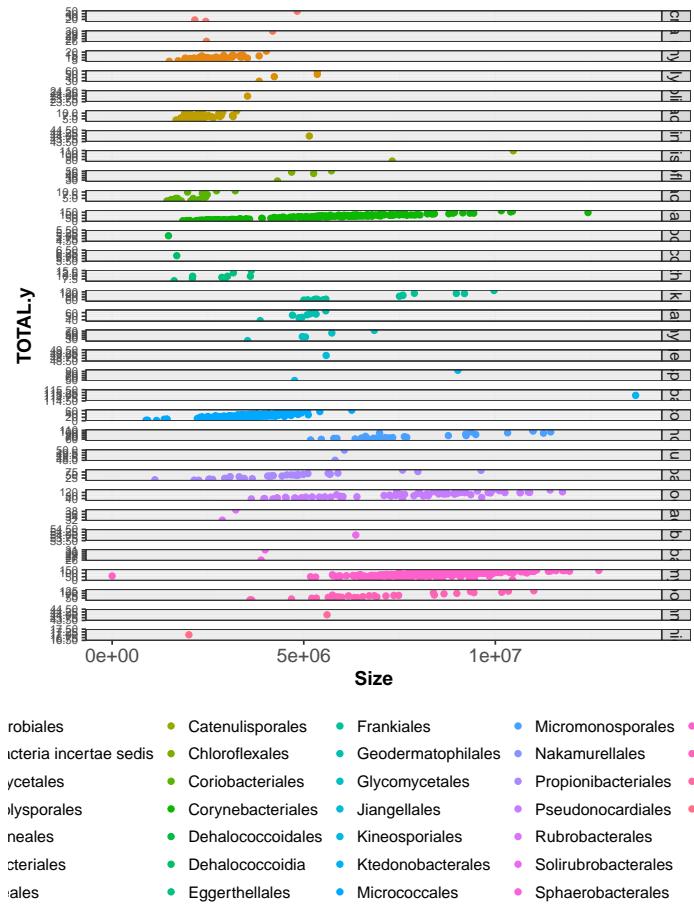


Figure 4.5: Correlation between Actinos genome size and antismash Natural products detection grided by Order

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

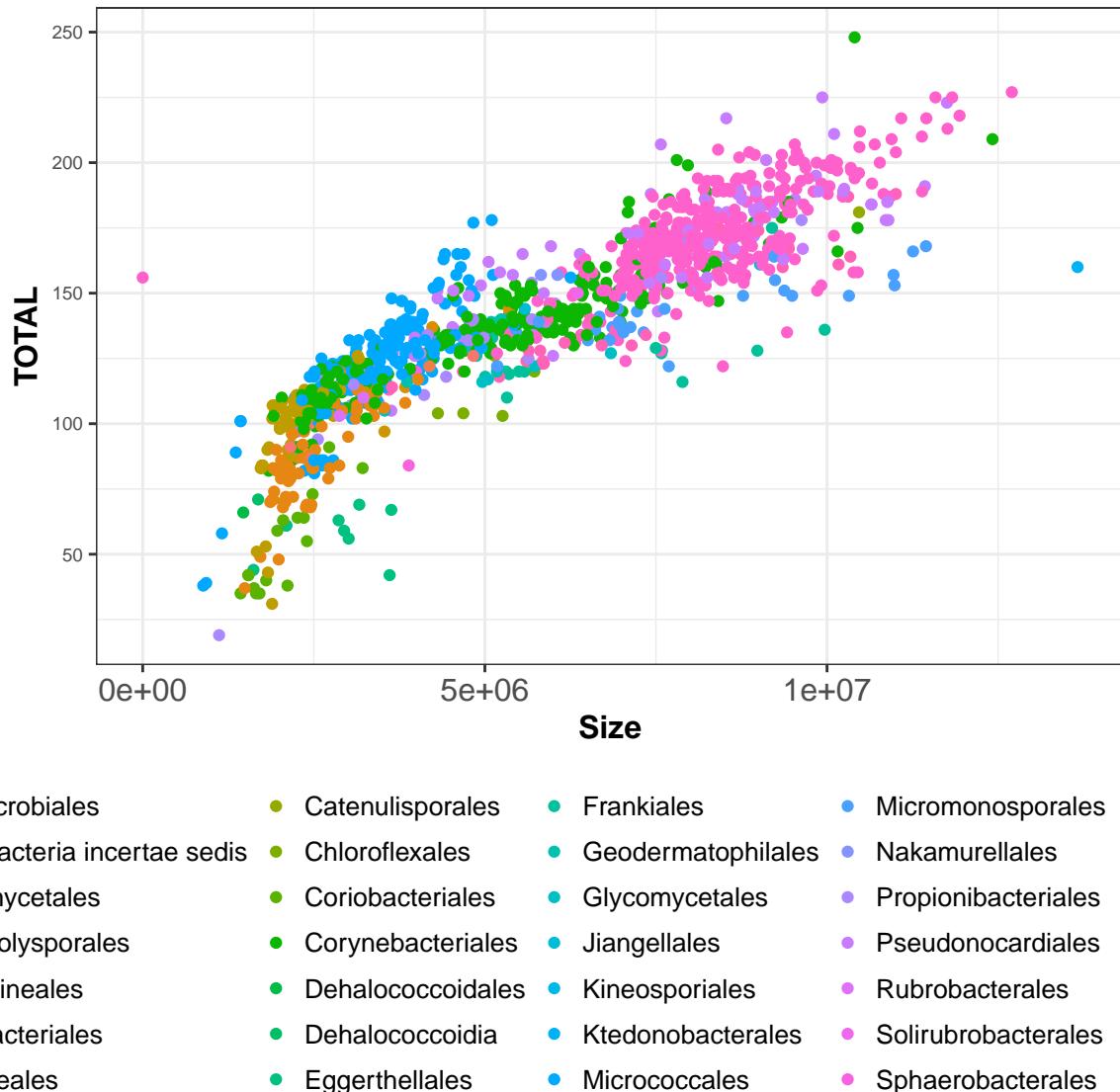


Figure 4.6: Correlation between Actinos genome size and central pathway expansions

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

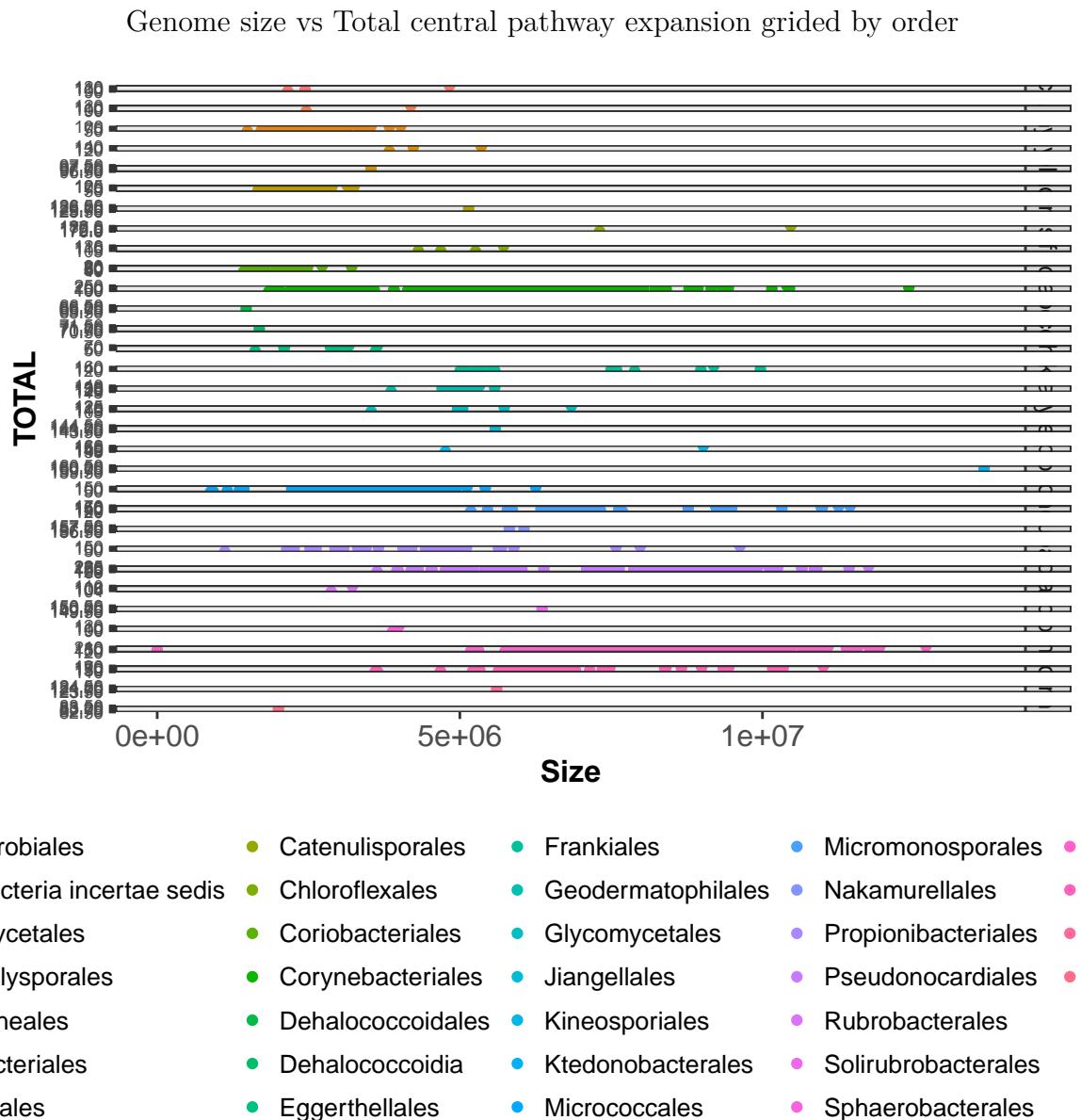


Figure 4.7: Correlation between Actinos genome size and central pathway expansions grided by order

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

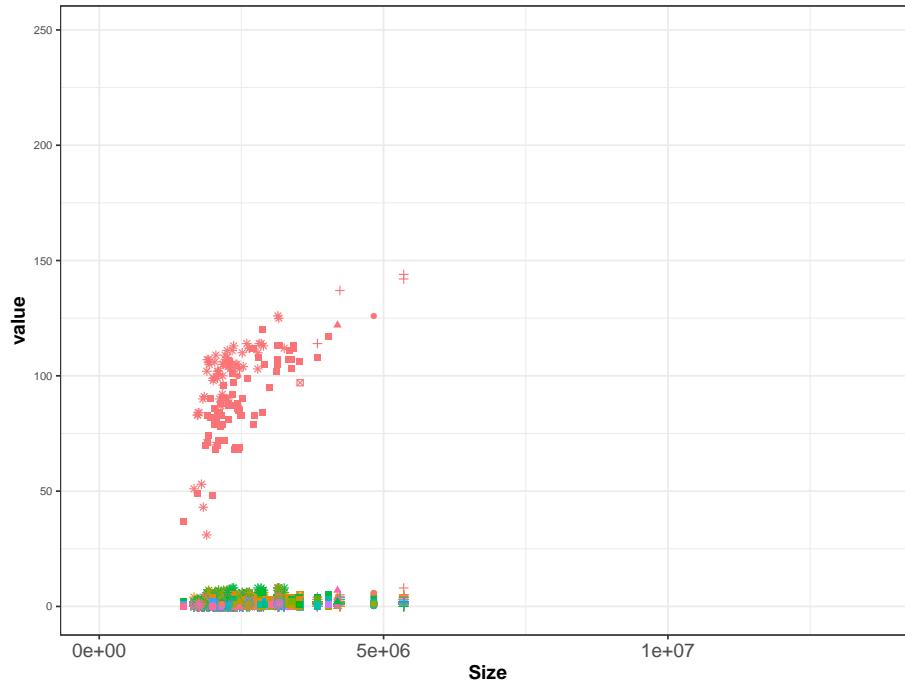


Figure 4.8: Correlation between Actinos Genome size vs Total central pathway expansion coloured by metabolic Family

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 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

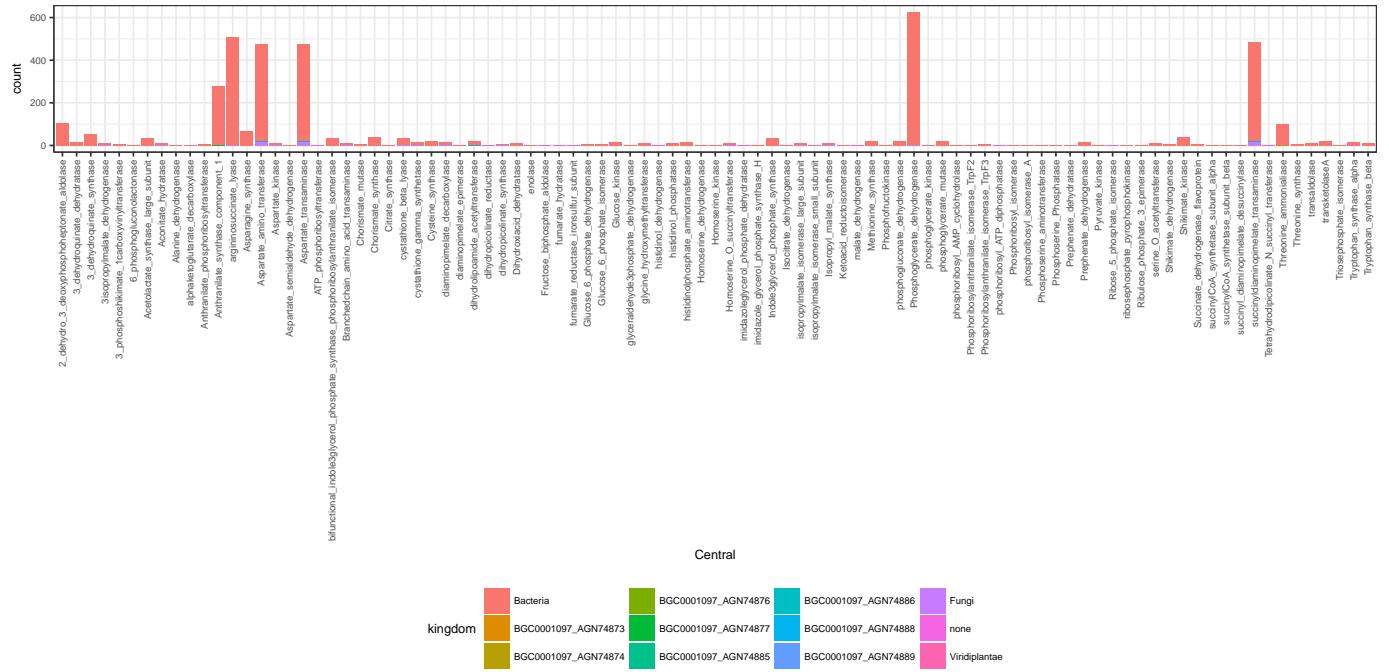


Figure 4.9: Actinos Recruitmens on central families coloured by kingdom

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

## Recruitments after central pathways expansions colourd by taxonomy

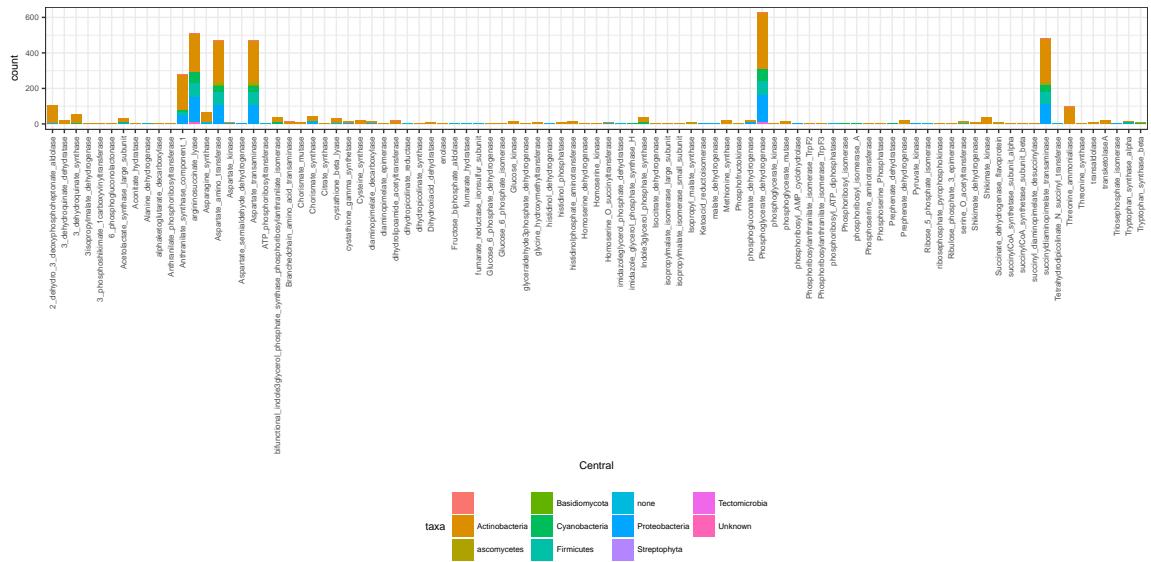


Figure 4.10: Actinos Recruitmens on central families coloured by taxonomy

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

## 4.5 Actinos AntiSMASH

## Taxonomical diversity on Actinosbacteria Data

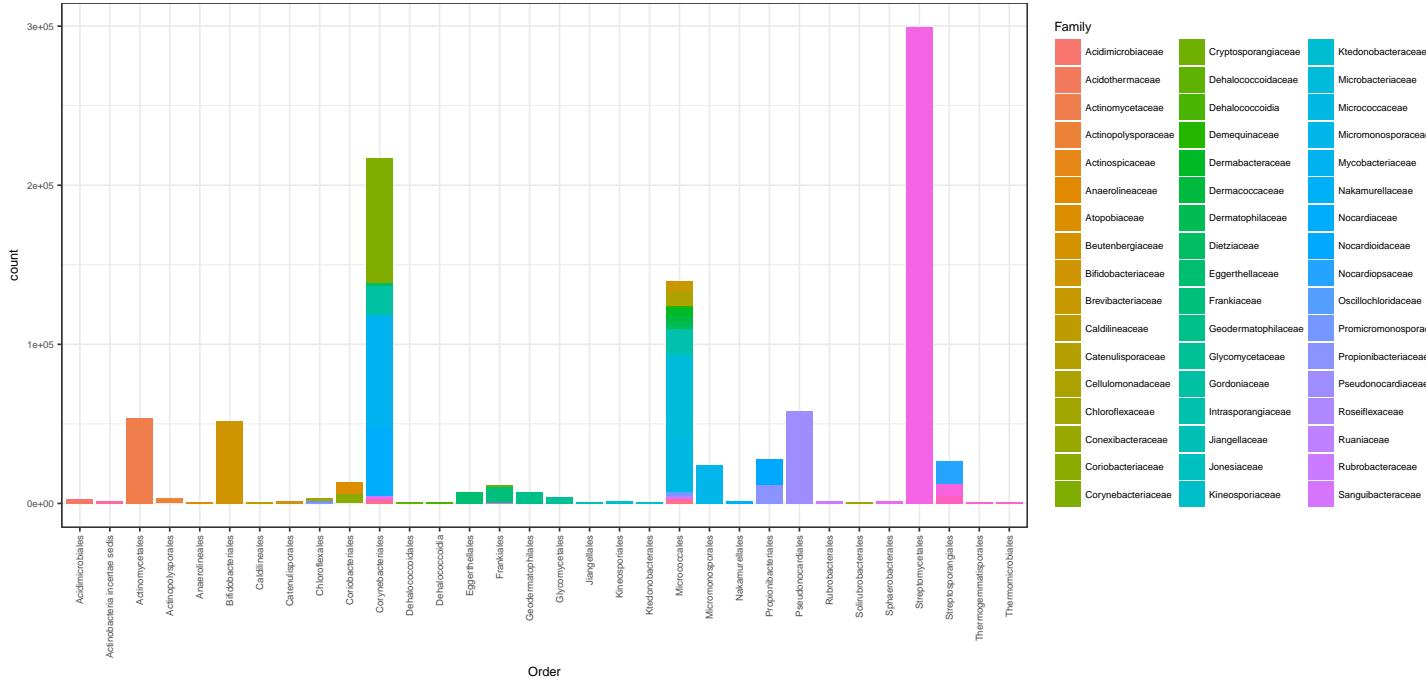


Figure 4.11: Actinos Diversity

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

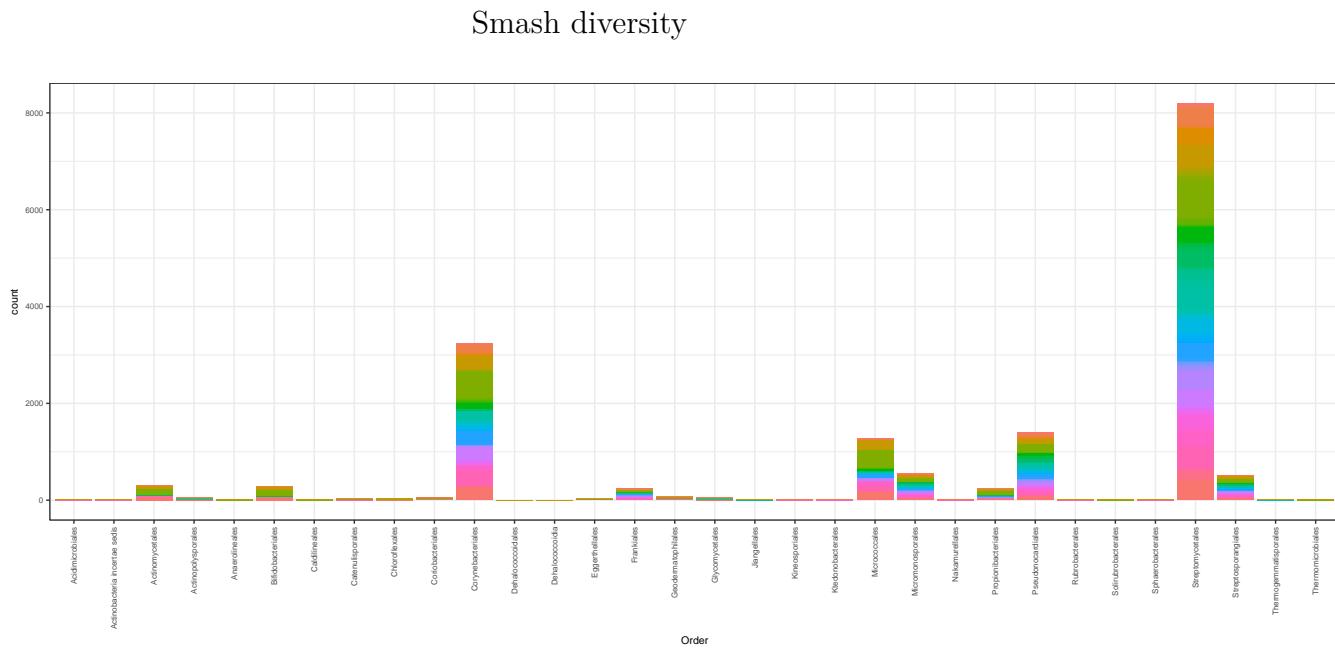


Figure 4.12: Actinos Smash Taxonomical Diversity

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

#### 4.5.1 AntisSMASH 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

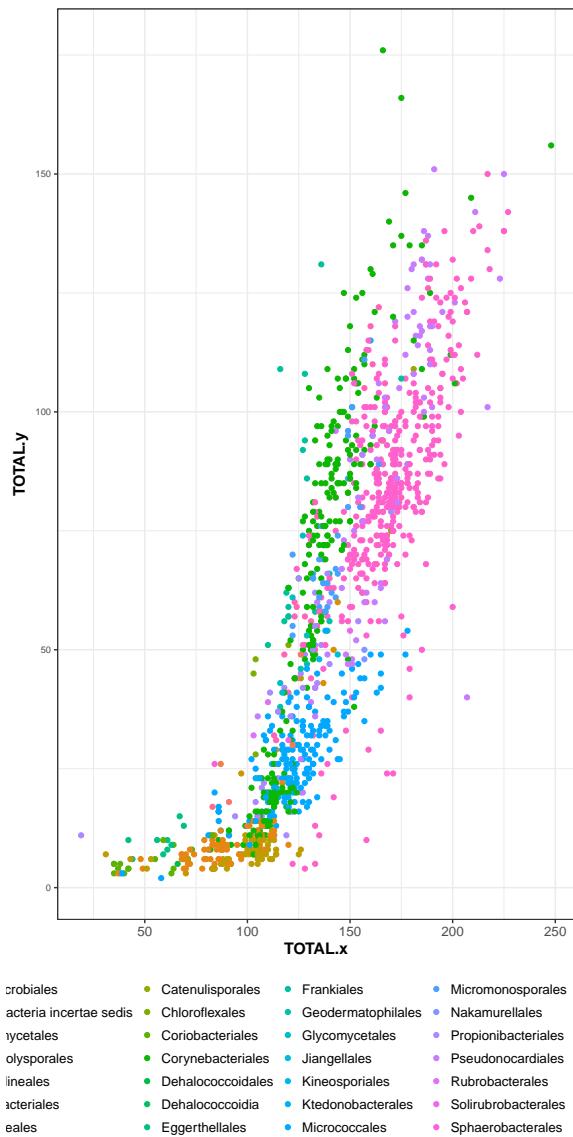


Figure 4.13: Correlation between Actinos central pathway expansions and antismash Natural products detection

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

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

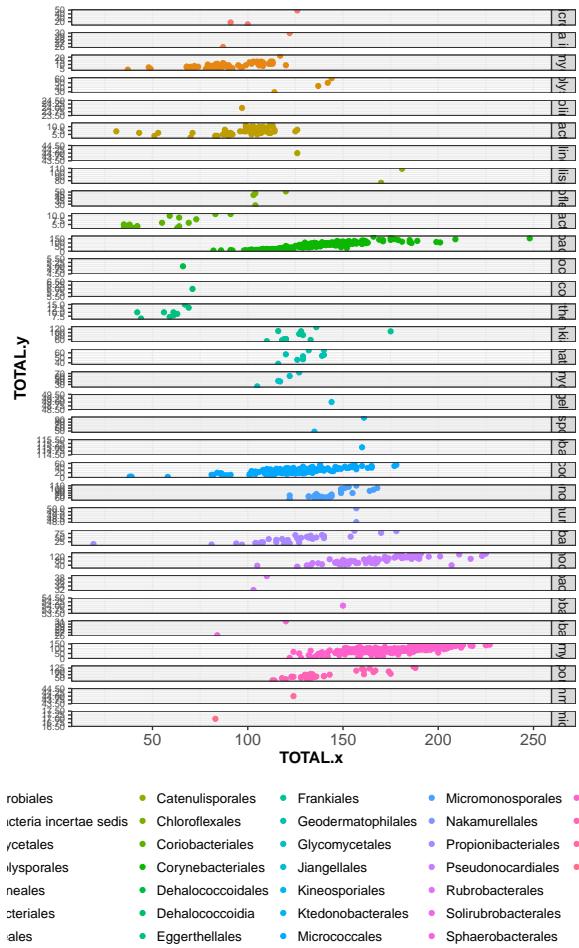


Figure 4.14: Correlation between Actinos central pathway expansions and antismash NP's clusters splitted by order plot

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

## Family

- |                       |                       |                         |               |
|-----------------------|-----------------------|-------------------------|---------------|
| • Acidimicrobiaceae   | • Cryptosporangiaceae | • Ktedonobacteraceae    | • Segnilipar  |
| • Acidothermaceae     | • Dehalococcoidaceae  | • Microbacteriaceae     | • Sphaerob    |
| • Actinomycetaceae    | • Dehalococcoidia     | • Micrococcaceae        | • Sporichthy  |
| • Actinopolysporaceae | • Demequinaceae       | • Micromonosporaceae    | • Streptomy   |
| • Actinospicaceae     | • Dermabacteraceae    | • Mycobacteriaceae      | • Streptospc  |
| • Anaerolineaceae     | • Dermacoccaceae      | • Nakamurellaceae       | • Thermoge    |
| • Atopobiaceae        | • Dermatophilaceae    | • Nocardiaceae          | • Thermomii   |
| • Beutenbergiaceae    | • Dietziaceae         | • Nocardioidaceae       | • Thermomc    |
| • Bifidobacteriaceae  | • Eggerthellaceae     | • Nocardiopsaceae       | • Timonella   |
| • Brevibacteriaceae   | • Frankiaceae         | • Oscillochloridaceae   | • Tsukamur    |
| • Caldilineaceae      | • Geodermatophilaceae | • Promicromonosporaceae | • Unclassifie |
| • Catenulisporaceae   | • Glycomycetaceae     | • Propionibacteriaceae  | • Unclassifie |
| • Cellulomonadaceae   | • Gordoniaceae        | • Pseudonocardiaceae    | • Unclassifie |
| • Chloroflexaceae     | • Intrasporangiaceae  | • Roseiflexaceae        | • Williamsia  |
| • Conexibacteraceae   | • Jiangellaceae       | • Ruaniaceae            |               |
| • Coriobacteriaceae   | • Jonesiaceae         | • Rubrobacteraceae      |               |
| • Corynebacteriaceae  | • Kineosporiaceae     | • Sanguibacteraceae     |               |

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

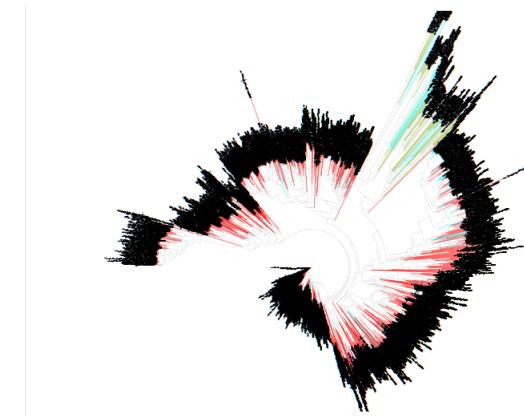


Figure 4.16: Enolase EvoMiningtree

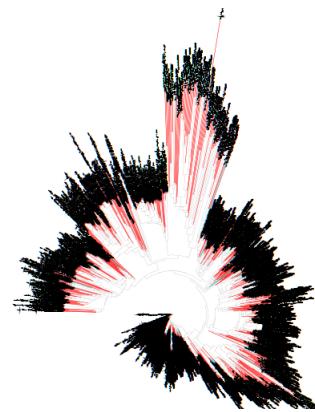


Figure 4.17: Phosphoribosyl isomerase EvoMiningtree

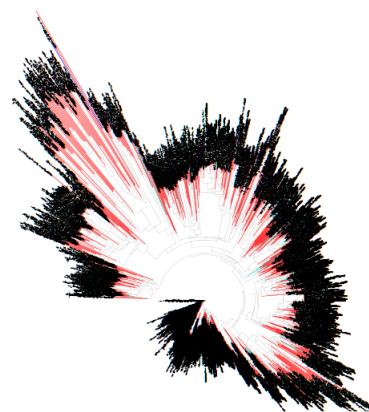


Figure 4.18: Phosphoribosyl isomerase A EvoMiningtree

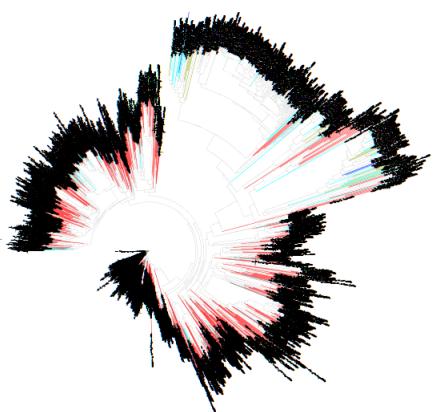


Figure 4.19: phosphoshikimate carboxyvinyltransferase EvoMiningtree

# Chapter 5

## Cyanobacteria EvoMining Results

Cyanobacteria phylum {Referencia}

Cyanobacteria is a photosynthetic phylum that inhabits a broad range of habitats. The broad adaptive potential is on part driven by gene-family enlargement [128] 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 [129] Example of a PriA cluster toxins[94]

Fossil record situates Cyanobacteria [129] Molecular record and metabolic properties at [132]

### 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.

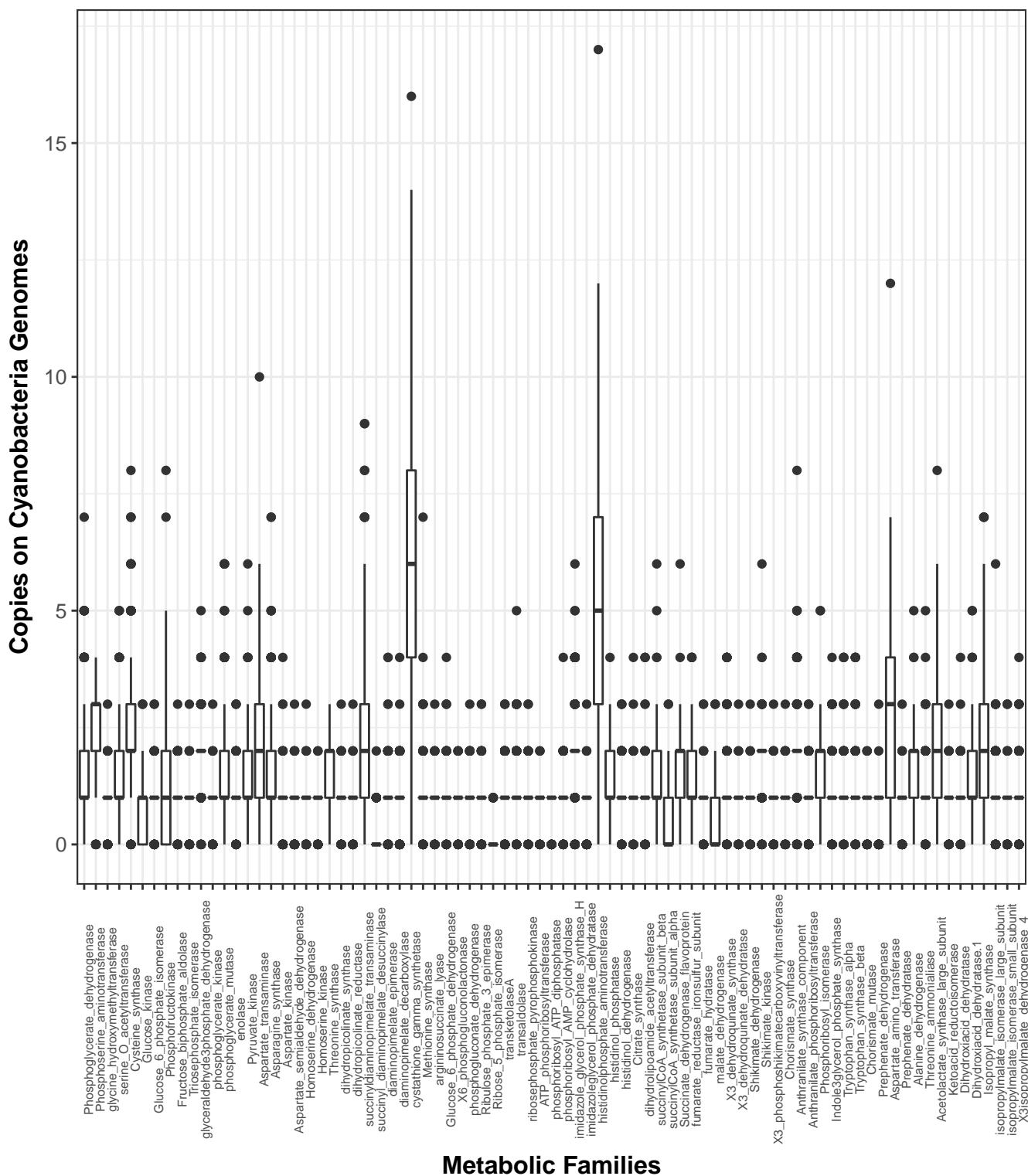


Figure 5.1: Expansions Boxplot

## 5.2 Central pathway expansions

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

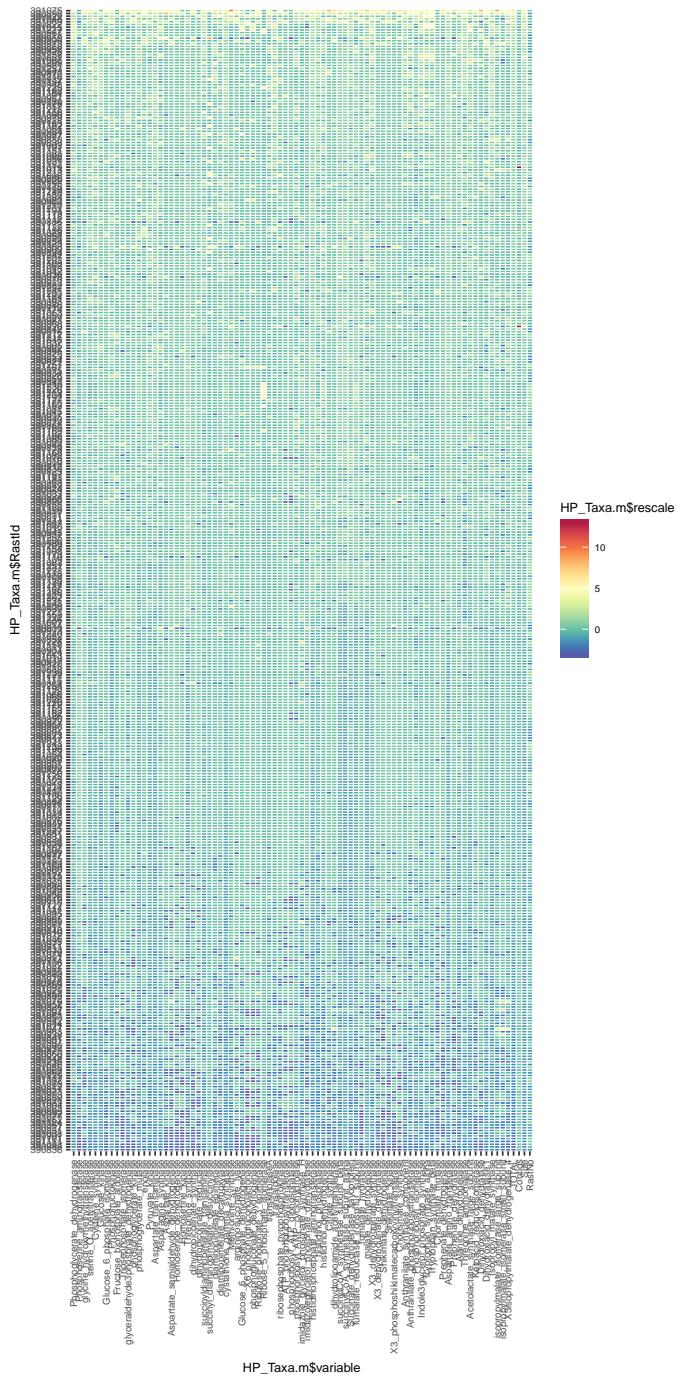


Figure 5.2: Cyanobacterial Heatplot

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

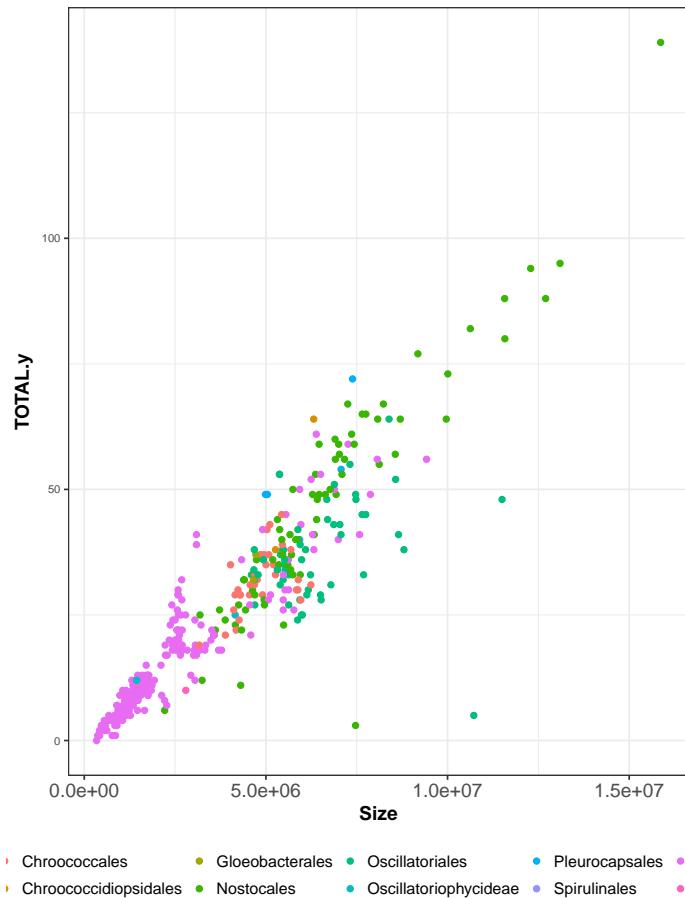


Figure 5.3: Correlation between genome size and antismash Natural products detection colored by Order

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

Genome size vs Total antismash cluster detected splitted by order

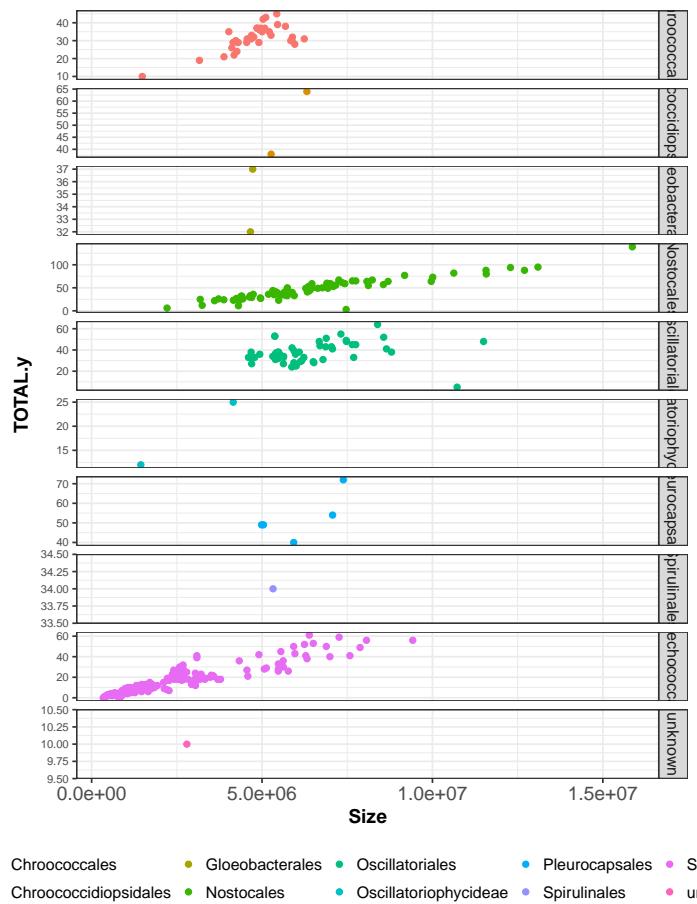


Figure 5.4: Correlation between genome size and antismash Natural products detection grided by Order

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

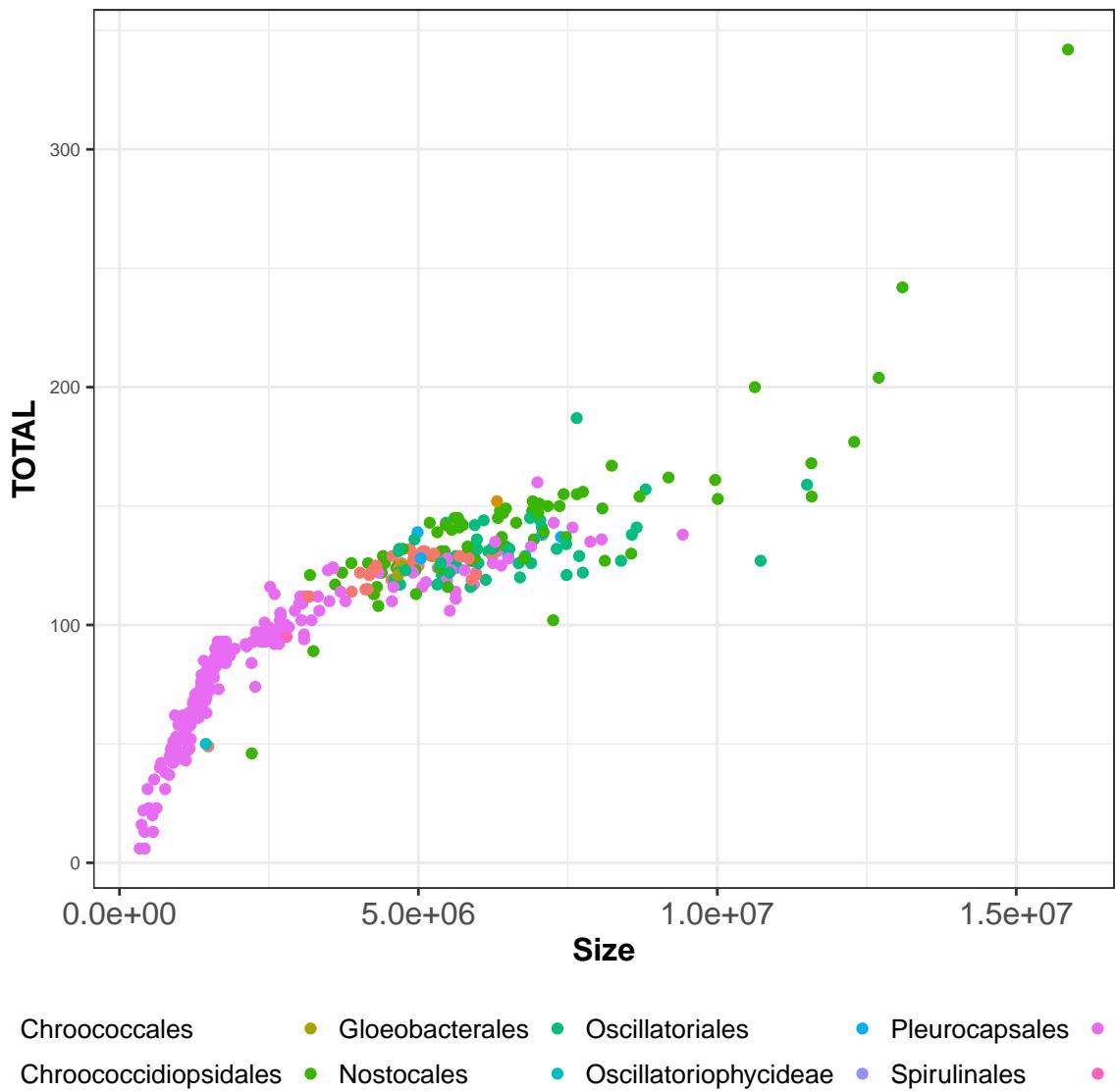


Figure 5.5: Correlation between genome size and central pathway expansions

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

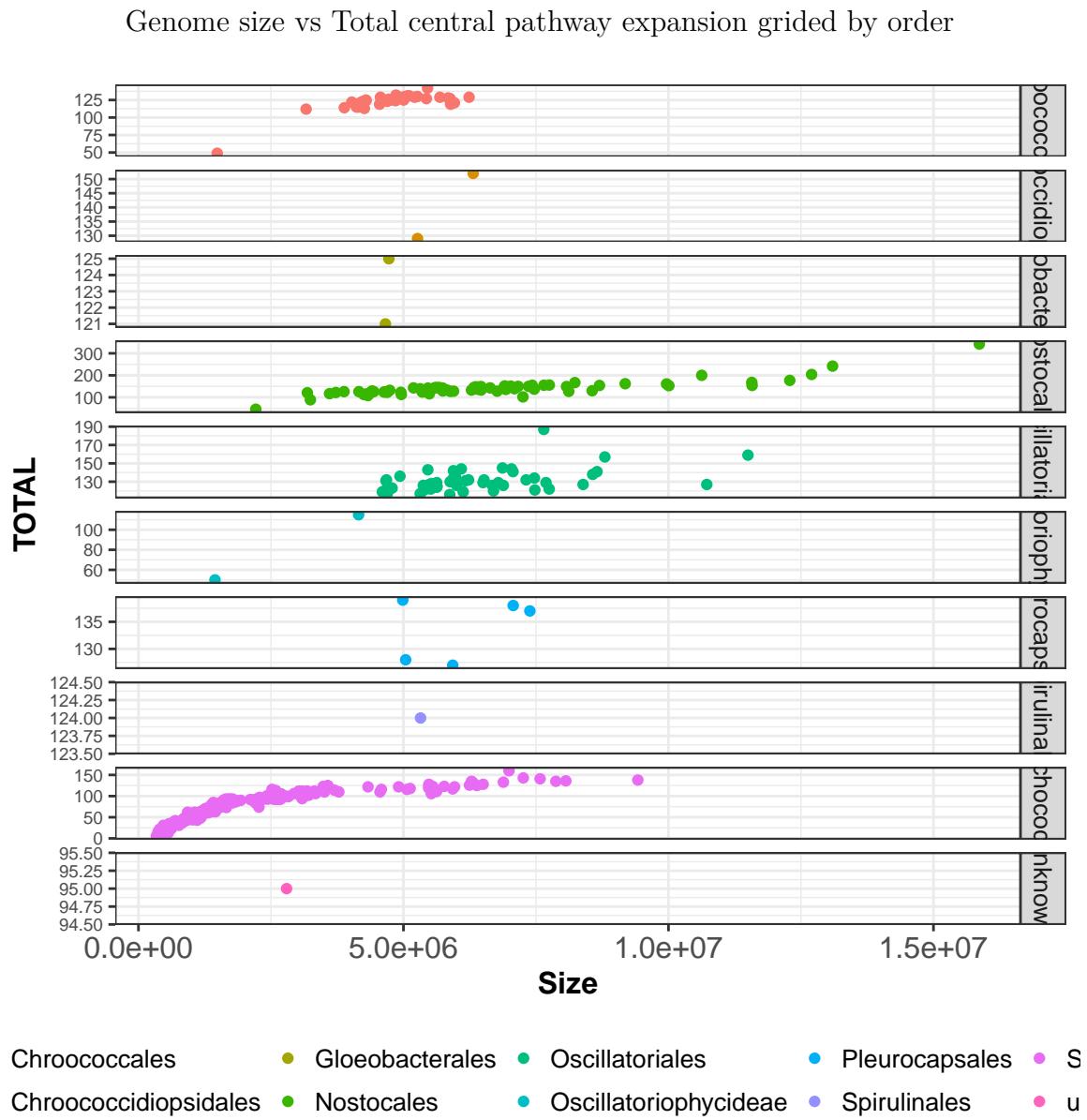


Figure 5.6: Correlation between genome size and central pathway expansions grided by order

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

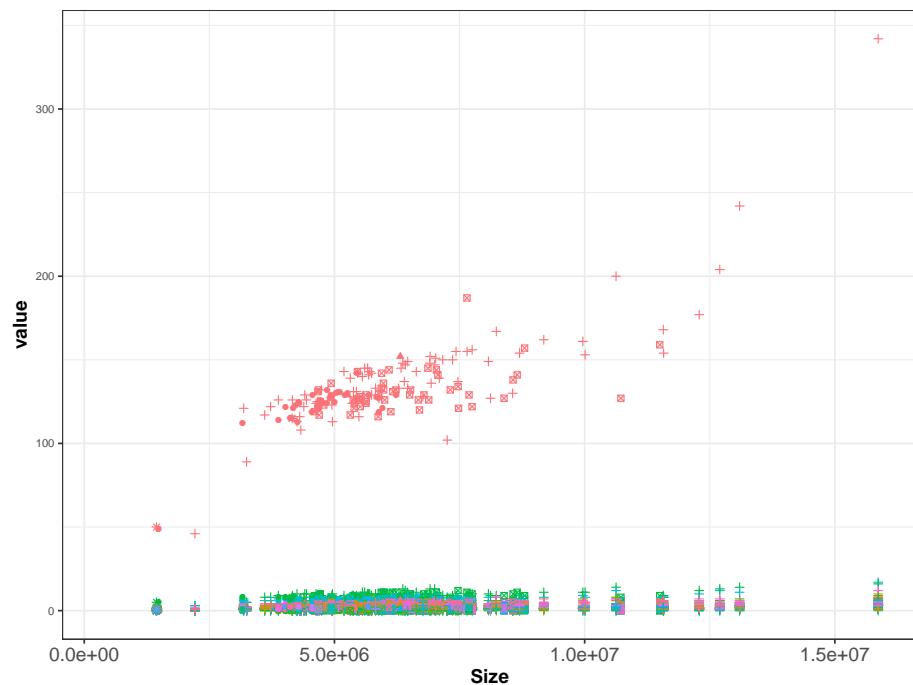


Figure 5.7: Correlation between Genome size vs Total central pathway expansion coloured by metabolic Family

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 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

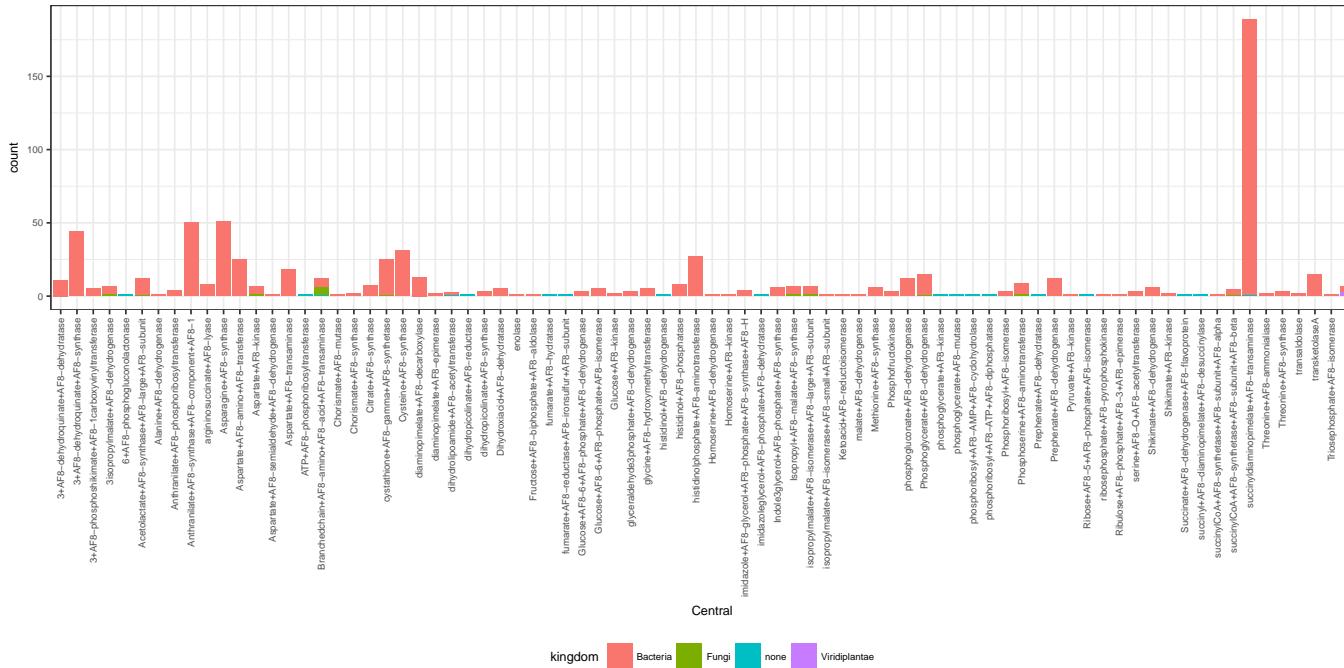


Figure 5.8: Recruitmens on central families coloured by kingdom

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

## Recruitments after central pathways expansions colourd by taxonomy

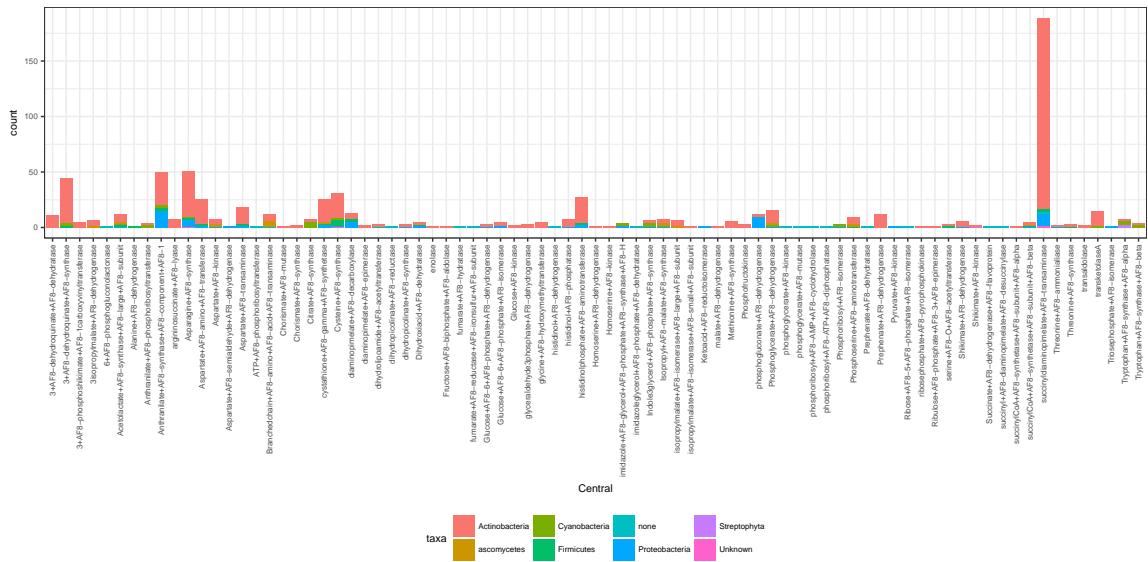


Figure 5.9: Recruitmens on central families coloured by taxonomy

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

## 5.5 Cyanobacterias AntiSMASH

Taxonomical diversity on Cyanobacteria Data

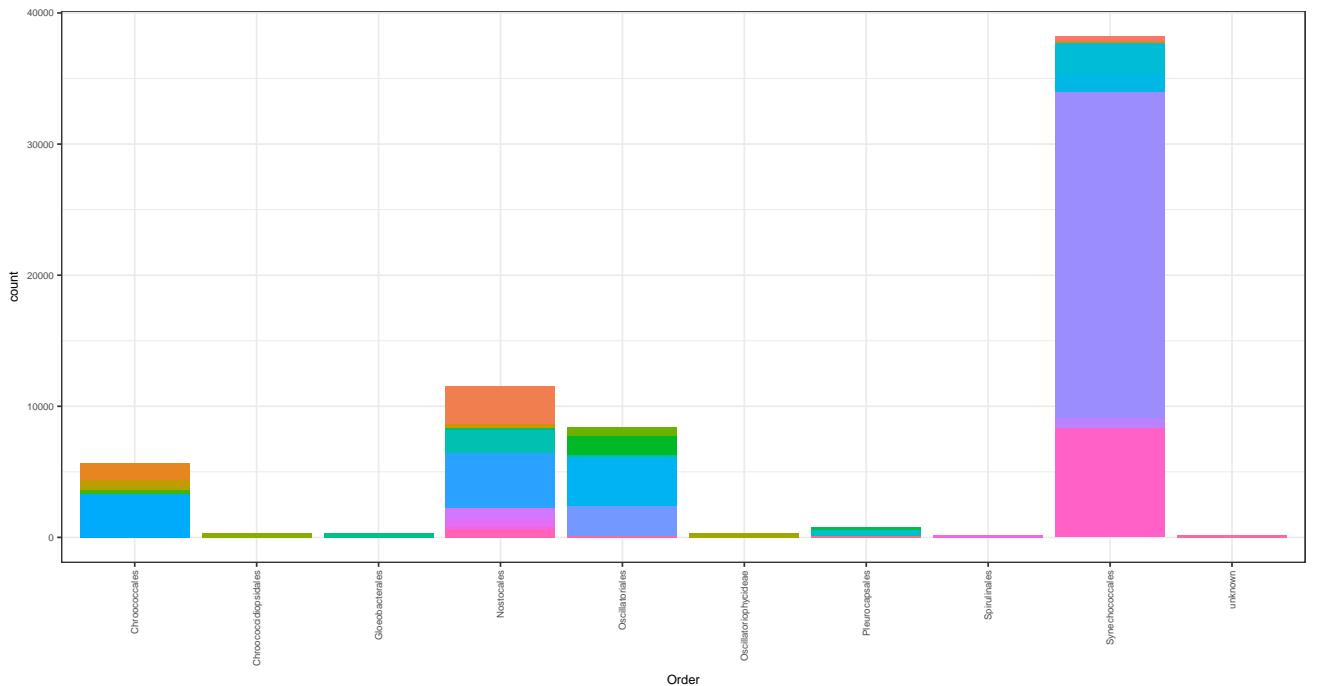


Figure 5.10: Diversity

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

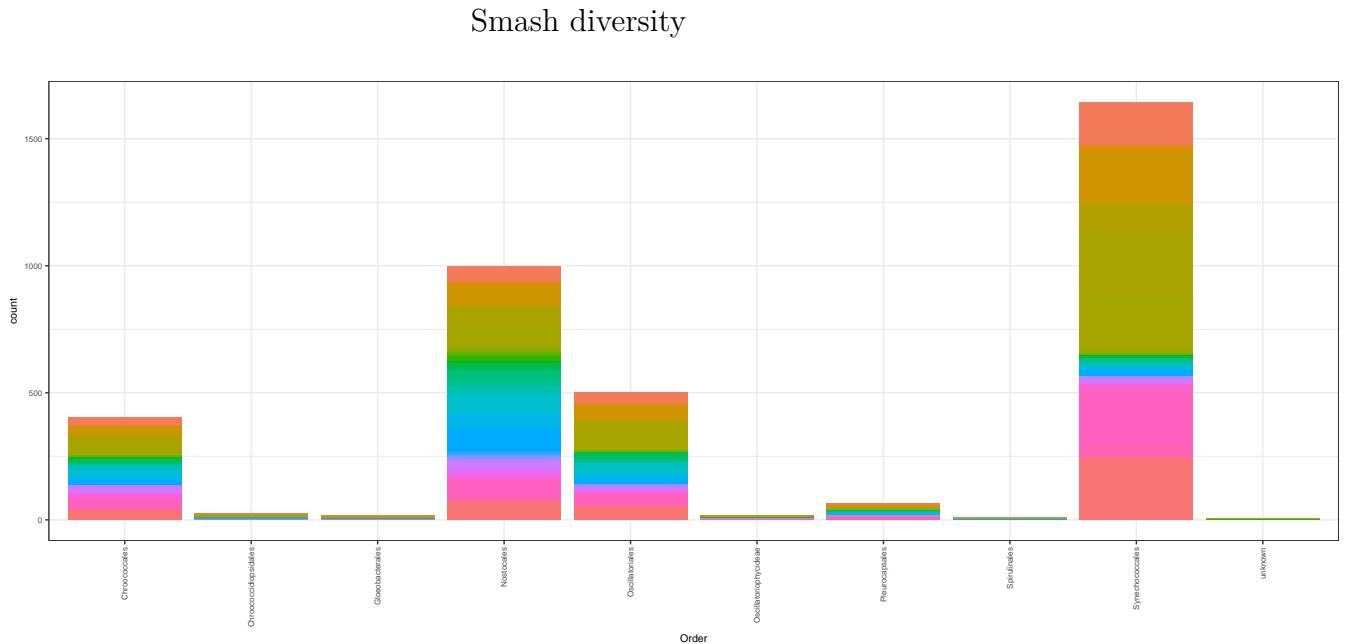


Figure 5.11: Smash

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

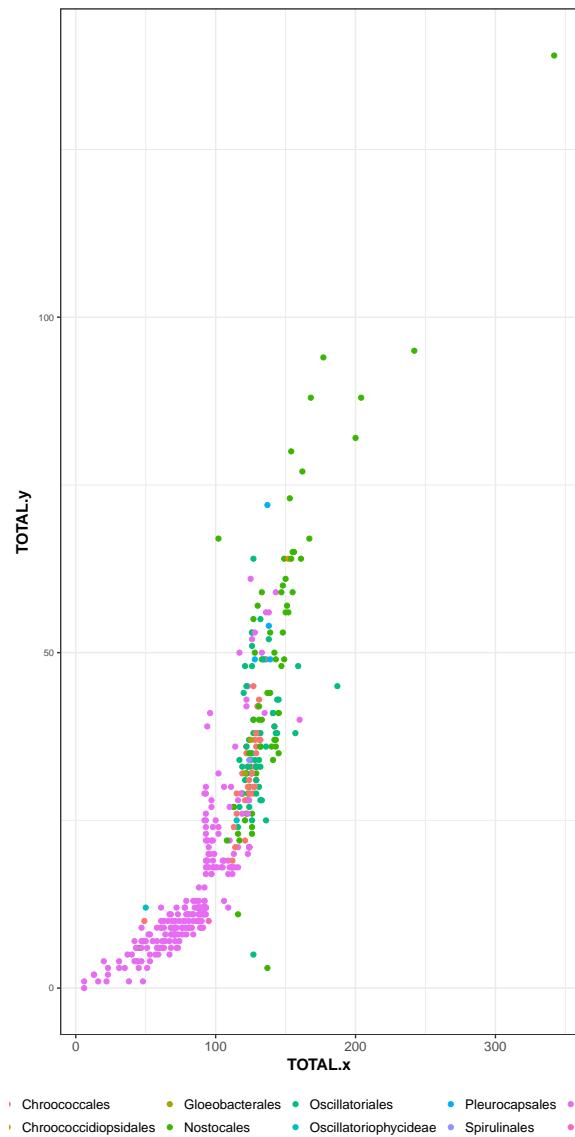


Figure 5.12: Correlation between central pathway expansions and anti-smash Natural products detection

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

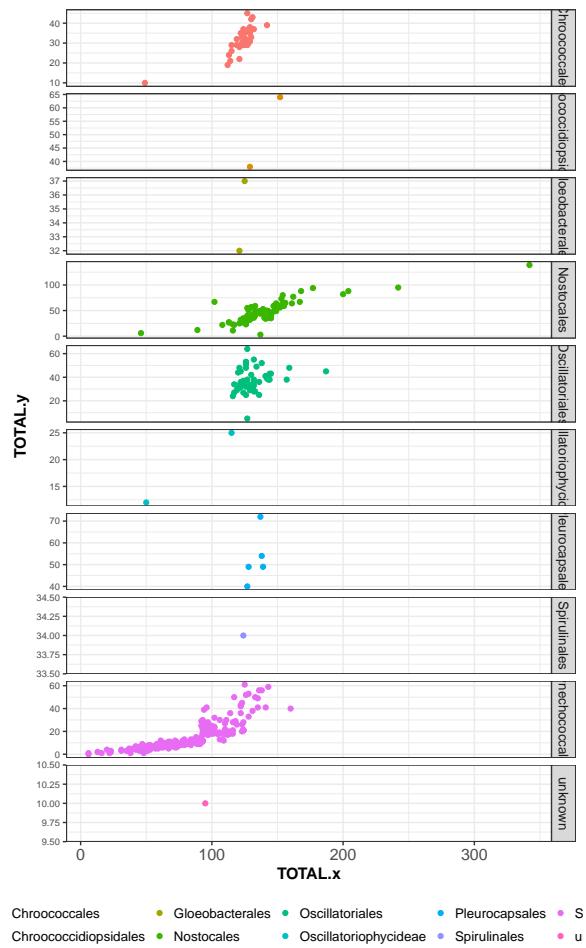


Figure 5.13: Correlation between central pathway expansions and anti-smash Natural products detection

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

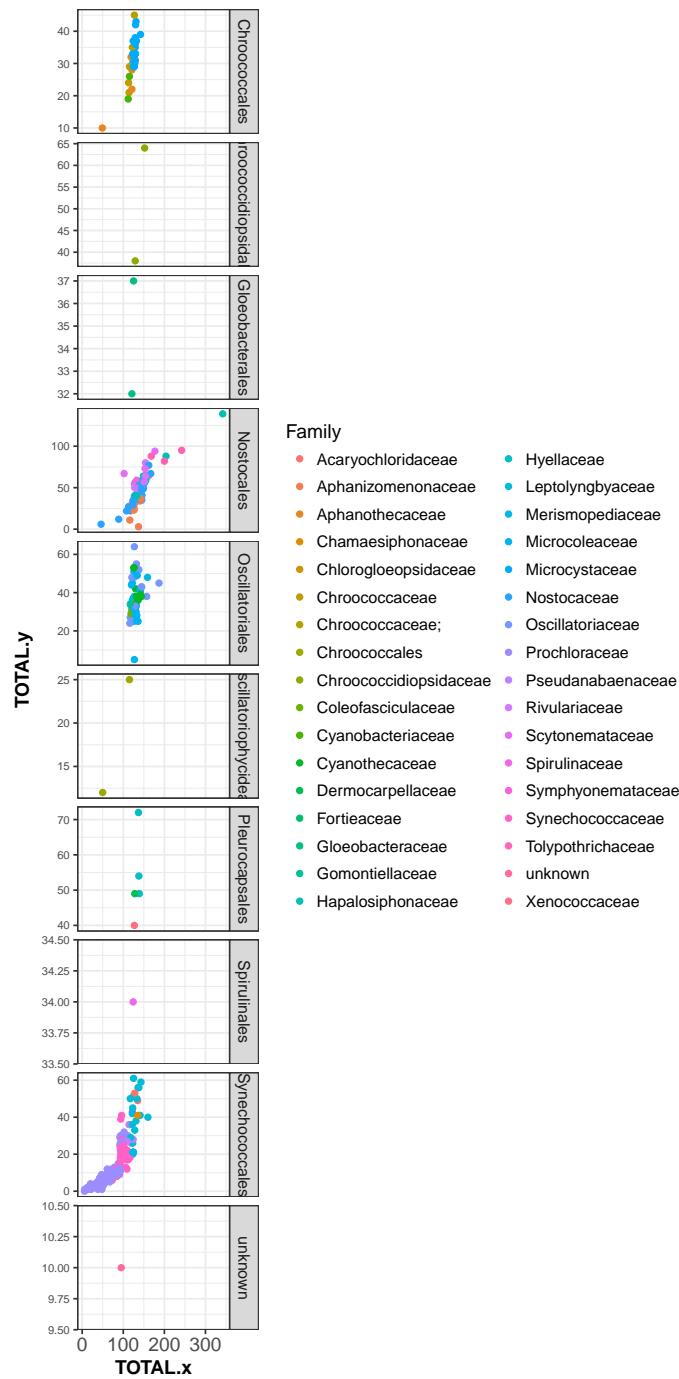


Figure 5.14: Natural products by family

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

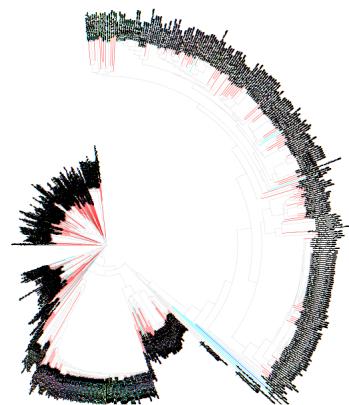


Figure 5.15: Phosphoribosyl isomerase EvoMiningtree

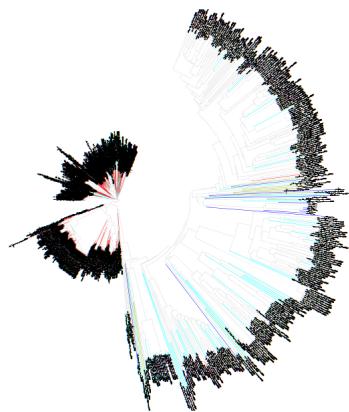


Figure 5.16: Phosphoglycerate dehydrogenase EvoMiningtree

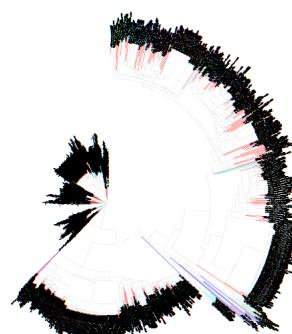


Figure 5.17: Phosphoserine aminotransferase EvoMiningtree

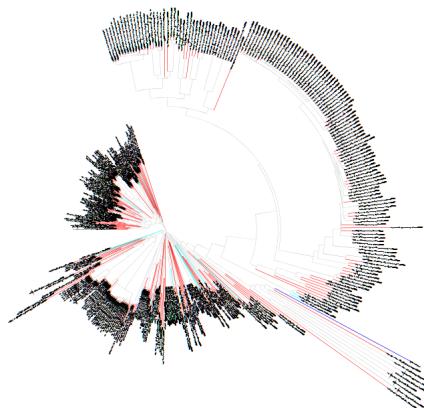


Figure 5.18: Triosephosphate isomerase EvoMiningtree

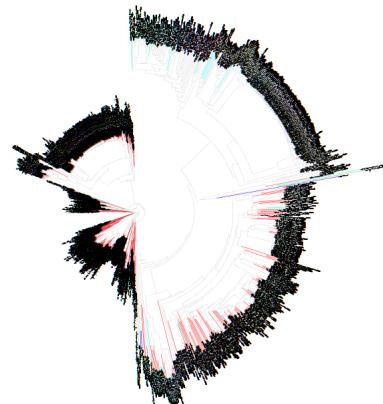


Figure 5.19: glyceraldehyde3phosphate dehydrogenase EvoMiningtree

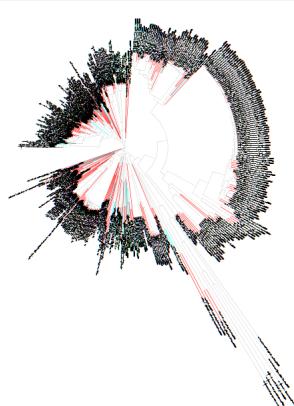


Figure 5.20: phosphoglycerate kinase EvoMiningtree

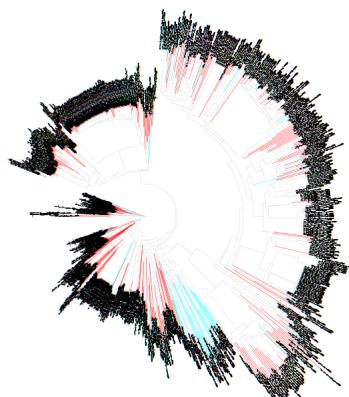


Figure 5.21: phosphoglycerate mutaseEvoMiningtree

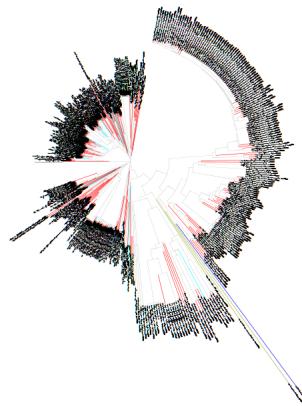


Figure 5.22: enolase EvoMiningtree

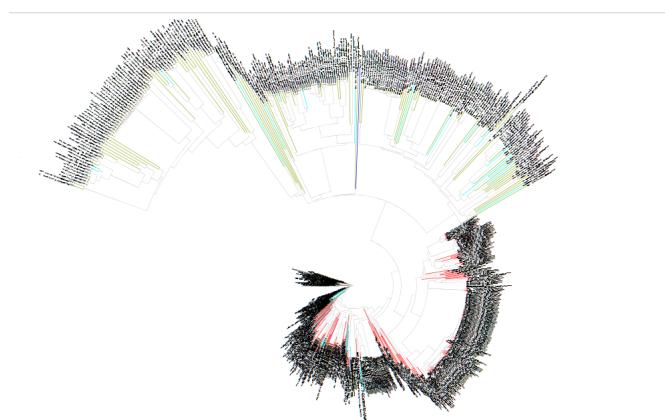


Figure 5.23: Pyruvate kinase EvoMiningtree

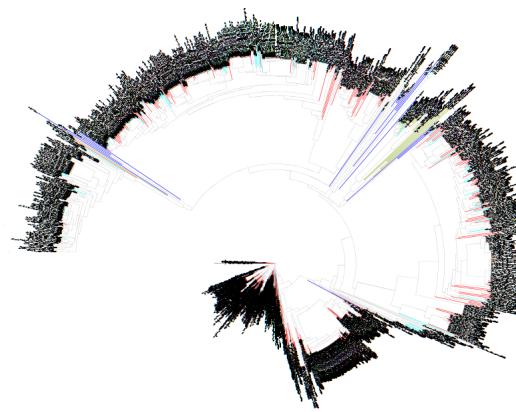


Figure 5.24: Aspartate transaminase EvoMiningtree

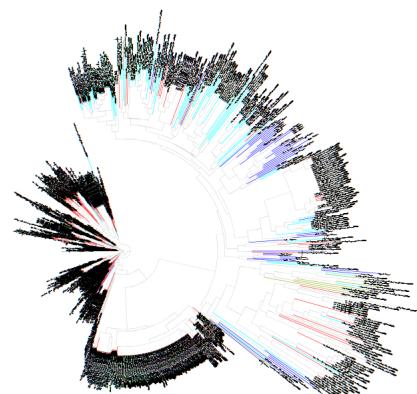


Figure 5.25: Asparagine synthase EvoMiningtree

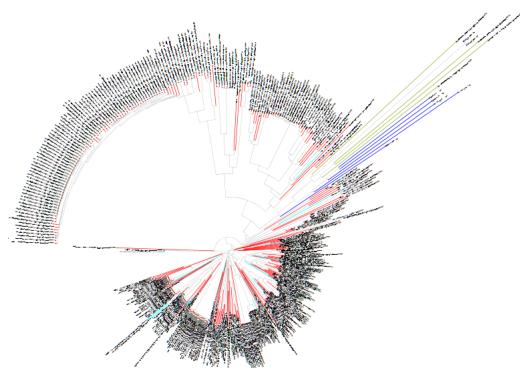


Figure 5.26: Aspartate kinase EvoMiningtree

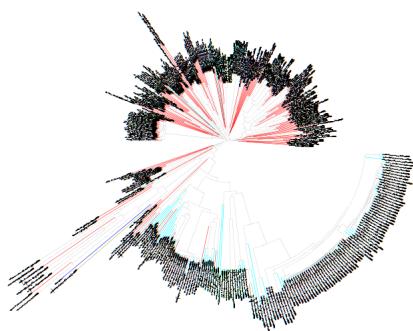


Figure 5.27: Aspartate semialdehyde dehydrogenase EvoMiningtree

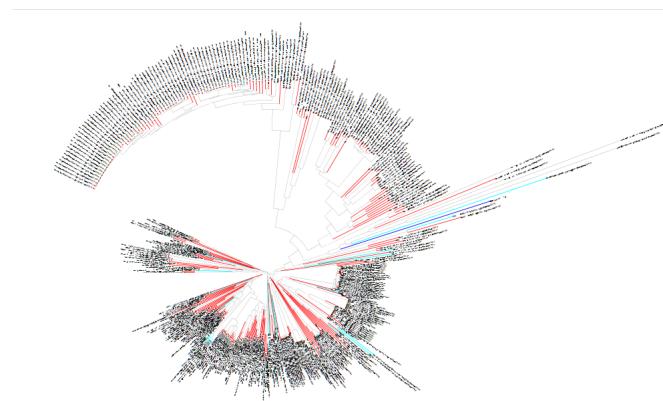


Figure 5.28: Homoserine dehydrogenase EvoMiningtree

# 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<sup>A</sup>T<sub>E</sub>X 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, Open source code on this document

### B.1 R markdown

Thanks to Rmardown Thesis  
Apendix one Useful docker commands  
-Create a new repository  
docker build . -t evomining  
docker push nselemevomining

### B.2 Docker

Restart docker and free all ports  
sudo service docker restart

list containers  
docker ps -a

ssh or bash into a running docker container  
sudo docker exec -i -t romantic\_brahmagupta /bin/bash docker exec -it  
<mycontainer> bash

Stop all containers  
docker rm \$(docker ps -a -q)

Remove stopped containers  
docker rm \$(docker ps -q -f status=exited)

Remove all images  
docker rmi \$(docker images -q)

```

uninstall docker from ubuntu (Fresh start)
    sudo apt-get purge docker-engine
    sudo apt-get autoremove --purge docker-engine
rm -rf /var/lib/docker # This deletes all images, containers, and volumes

Run Evomining container using nselem/newevomining image
    docker run -i -t -v
/home/nelly/GIT/EvoMining/:/var/www/html/EvoMining/exchange -p 80:80
    nselem/newevomining /bin/bash

Start evomining inside this container
    perl startevomining

Vizualice a tree
http://10.10.100.234/EvoMining/cgi-bin/color_tree.pl?9&&/var/www/html/EvoMining/exchange
file 9.new must be on folder volume CyanosBBH_MiBIG_DB.faa_CYANOS

Find a perl module
perl -MList::Util -e'print $_ . " => " . $INC{$_} . "\n" for keys
%INC' EvoMining notes
Gblocks only runs inside folder /var/www/html/EvoMining

```

### B.3 Git

```

git add --all
git commit -m "Some message"
git push -u origin master
git clone

```

### B.4 Connect GitHub and DockerHub

automated builds The Dockerfile is available to anyone with access to your Docker Hub repository. Your repository is kept up-to-date with code changes automatically.

### B.5 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/>



# References

1. Angel E. Interactive computer graphics : A top-down approach with OpenGL. Boston, MA: Addison Wesley Longman; 2000.
2. Angel E. Batch-file computer graphics : A bottom-up approach with quickTime. Boston, MA: Wesley Addison Longman; 2001.
3. Angel E. Test second book by angel. Boston, MA: Wesley Addison Longman; 2001.
4. Khersonsky O, Tawfik DS. Enzyme promiscuity: A mechanistic and evolutionary perspective. *Annual Review of Biochemistry*. 2010;79: 471–505.  
doi:10.1146/annurev-biochem-030409-143718
5. Copley SD. Enzymes with extra talents: Moonlighting functions and catalytic promiscuity. *Current Opinion in Chemical Biology*. 2003;7: 265–272.  
doi:10.1016/S1367-5931(03)00032-2
6. Hult K, Berglund P. Enzyme promiscuity: Mechanism and applications. *Trends in Biotechnology*. 2007;25: 231–238. doi:10.1016/j.tibtech.2007.03.002
7. O'Brien PJ, Herschlag D. Catalytic promiscuity and the evolution of new enzymatic activities. *Chemistry & Biology*. 1999;6: R91–R105.  
doi:10.1016/S1074-5521(99)80033-7
8. Barona Gómez F, Hodgson DA. Occurrence of a putative ancient like isomerase involved in histidine and tryptophan biosynthesis. *EMBO reports*. 2003;4: 296–300. doi:10.1038/sj.embor.embor771
9. Risso VA, Gavira JA, Gaucher EA, Sanchez Ruiz JM. Phenotypic comparisons of consensus variants versus laboratory resurrections of precambrian proteins. *Proteins: Structure, Function, and Bioinformatics*. 2014;82: 887–896.  
doi:10.1002/prot.24575
10. Kumari V, Shah S, Gupta MN. Preparation of Biodiesel by Lipase-Catalyzed Transesterification of High Free Fatty Acid Containing Oil from Madhuca indica. *Energy & Fuels*. 2007;21: 368–372. doi:10.1021/ef0602168
11. Li C, Henry CS, Jankowski MD, Ionita JA, Hatzimanikatis V, Broadbelt LJ. Computational discovery of biochemical routes to specialty chemicals. *Chemical*

- Engineering Science. 2004;59: 5051–5060. doi:10.1016/j.ces.2004.09.021
12. Glasner ME, Gerlt JA, Babbitt PC. Evolution of enzyme superfamilies. Current Opinion in Chemical Biology. 2006;10: 492–497. doi:10.1016/j.cbpa.2006.08.012
13. Baier F, Copp JN, Tokuriki N. Evolution of Enzyme Superfamilies: Comprehensive Exploration of Sequence–Function Relationships. Biochemistry. 2016;55: 6375–6388. doi:10.1021/acs.biochem.6b00723
14. Bloom JD, Romero PA, Lu Z, Arnold FH. Neutral genetic drift can alter promiscuous protein functions, potentially aiding functional evolution. Biology Direct. 2007;2: 17. doi:10.1186/1745-6150-2-17
15. Nath A, Atkins WM. A Quantitative Index of Substrate Promiscuity. Biochemistry. 2008;47: 157–166. doi:10.1021/bi701448p
16. Zou T, Risso VA, Gavira JA, Sanchez-Ruiz JM, Ozkan SB. Evolution of Conformational Dynamics Determines the Conversion of a Promiscuous Generalist into a Specialist Enzyme. Molecular Biology and Evolution. 2015;32: 132–143. doi:10.1093/molbev/msu281
17. Firn RD, Jones CG. A Darwinian view of metabolism: Molecular properties determine fitness. Journal of Experimental Botany. 2009;60: 719–726. doi:10.1093/jxb/erp002
18. Jia B, Cheong G-W, Zhang S. Multifunctional enzymes in archaea: Promiscuity and moonlight. Extremophiles. 2013;17: 193–203. doi:10.1007/s00792-012-0509-1
19. Aharoni A, Gaidukov L, Khersonsky O, Gould SM, Roodveldt C, Tawfik DS. The 'evolvability' of promiscuous protein functions. Nature Genetics. 2005;37: 73–76. doi:10.1038/ng1482
20. Jensen. Enzyme Recruitment in Evolution of New Function. Annual Review of Microbiology. 1976;30: 409–425. doi:10.1146/annurev.mi.30.100176.002205
21. Pandya C, Farelli JD, Dunaway-Mariano D, Allen KN. Enzyme Promiscuity: Engine of Evolutionary Innovation. Journal of Biological Chemistry. 2014;289: 30229–30236. doi:10.1074/jbc.R114.572990
22. Dean AM, Thornton JW. Mechanistic approaches to the study of evolution. Nature reviews Genetics. 2007;8: 675–688. doi:10.1038/nrg2160
23. Nobeli I, Favia AD, Thornton JM. Protein promiscuity and its implications for biotechnology. Nature Biotechnology. 2009;27: 157–167. doi:10.1038/nbt1519
24. Hopkins AL. Drug discovery: Predicting promiscuity. Nature. 2009;462: 167–168. doi:10.1038/462167a
25. Nath A, Zientek MA, Burke BJ, Jiang Y, Atkins WM. Quantifying and Predicting the Promiscuity and Isoform Specificity of Small-Molecule Cytochrome P450 Inhibitors. Drug Metabolism and Disposition. 2010;38: 2195–2203.

- doi:10.1124/dmd.110.034645
26. Eichborn J von, Murgueitio MS, Dunkel M, Koerner S, Bourne PE, Preissner R. PROMISCUOUS: A database for network-based drug-repositioning. *Nucleic Acids Research*. 2011;39: D1060–D1066. doi:10.1093/nar/gkq1037
  27. Zhang W, Dourado DFAR, Fernandes PA, Ramos MJ, Mannervik B. Multidimensional epistasis and fitness landscapes in enzyme evolution. *Biochemical Journal*. 2012;445: 39–46. doi:10.1042/BJ20120136
  28. Sanchez-Ruiz JM. On promiscuity, changing environments and the possibility of replaying the molecular tape of life. *Biochemical Journal*. 2012;445: e1–e3. doi:10.1042/BJ20120806
  29. Martínez-Núñez MA, Rodríguez-Vázquez K, Pérez-Rueda E. The lifestyle of prokaryotic organisms influences the repertoire of promiscuous enzymes. *Proteins: Structure, Function, and Bioinformatics*. 2015;83: 1625–1631. doi:10.1002/prot.24847
  30. Patrick WM, Quandt EM, Swartzlander DB, Matsumura I. Multicopy Suppression Underpins Metabolic Evolvability. *Molecular Biology and Evolution*. 2007;24: 2716–2722. doi:10.1093/molbev/msm204
  31. Notebaart RA, Szappanos B, Kintses B, Pál F, Györke Á, Bogos B, et al. Network-level architecture and the evolutionary potential of underground metabolism. *Proceedings of the National Academy of Sciences*. 2014;111: 11762–11767. doi:10.1073/pnas.1406102111
  32. Linster CL, Van Schaftingen E, Hanson AD. Metabolite damage and its repair or pre-emption. *Nature Chemical Biology*. 2013;9: 72–80. doi:10.1038/nchembio.1141
  33. Khanal A, Yu McLoughlin S, Kershner JP, Copley SD. Differential Effects of a Mutation on the Normal and Promiscuous Activities of Orthologs: Implications for Natural and Directed Evolution. *Molecular Biology and Evolution*. 2015;32: 100–108. doi:10.1093/molbev/msu271
  34. Ma H-M, Zhou Q, Tang Y-M, Zhang Z, Chen Y-S, He H-Y, et al. Unconventional Origin and Hybrid System for Construction of Pyrrolopyrrole Moiety in Kosinostatin Biosynthesis. *Chemistry & Biology*. 2013;20: 796–805. doi:10.1016/j.chembiol.2013.04.013
  35. Adams NE, Thiaville JJ, Proestos J, Juárez-Vázquez AL, McCoy AJ, Barona-Gómez F, et al. Promiscuous and Adaptable Enzymes Fill “Holes” in the Tetrahydrofolate Pathway in Chlamydia Species. *mBio*. 2014;5. doi:10.1128/mBio.01378-14
  36. Soskine M, Tawfik DS. Mutational effects and the evolution of new protein functions. *Nature Reviews Genetics*. 2010;11: 572–582. doi:10.1038/nrg2808
  37. Halachev MR, Loman NJ, Pallen MJ. Calculating Orthologs in Bacteria and

- Archaea: A Divide and Conquer Approach. PLOS ONE. 2011;6: e28388. doi:10.1371/journal.pone.0028388
38. Kislyuk AO, Haegeman B, Bergman NH, Weitz JS. Genomic fluidity: An integrative view of gene diversity within microbial populations. BMC Genomics. 2011;12: 32. doi:10.1186/1471-2164-12-32
39. Pearson H. Prehistoric proteins: Raising the dead. Nature News. 2012;483: 390. doi:10.1038/483390a
40. Hughes AL. The Evolution of Functionally Novel Proteins after Gene Duplication. Proceedings of the Royal Society of London B: Biological Sciences. 1994;256: 119–124. doi:10.1098/rspb.1994.0058
41. Treangen TJ, Rocha EPC. Horizontal Transfer, Not Duplication, Drives the Expansion of Protein Families in Prokaryotes. PLOS Genetics. 2011;7: e1001284. doi:10.1371/journal.pgen.1001284
42. Overbeek R, Fonstein M, D’Souza M, Pusch GD, Maltsev N. The use of gene clusters to infer functional coupling. Proceedings of the National Academy of Sciences. 1999;96: 2896–2901. doi:10.1073/pnas.96.6.2896
43. Zhao S, Sakai A, Zhang X, Vetting MW, Kumar R, Hillerich B, et al. Prediction and characterization of enzymatic activities guided by sequence similarity and genome neighborhood networks. eLife. 2014;3: e03275. doi:10.7554/eLife.03275
44. Zhao S, Kumar R, Sakai A, Vetting MW, Wood BM, Brown S, et al. Discovery of new enzymes and metabolic pathways by using structure and genome context. Nature. 2013;502: 698–702. doi:10.1038/nature12576
45. Verdel-Aranda K, López-Cortina ST, Hodgson DA, Barona-Gómez F. Molecular annotation of ketol-acid reductoisomerase from Streptomyces reveals a novel amino acid biosynthesis interlock mediated by enzyme promiscuity. Microbial Biotechnology. 2015;8: 239–252. doi:10.1111/1751-7915.12175
46. Szklarczyk D, Franceschini A, Wyder S, Forslund K, Heller D, Huerta-Cepas J, et al. STRING v10: Protein–protein interaction networks, integrated over the tree of life. Nucleic Acids Research. 2015;43: D447–D452. doi:10.1093/nar/gku1003
47. Snel B, Lehmann G, Bork P, Huynen MA. STRING: A web-server to retrieve and display the repeatedly occurring neighbourhood of a gene. Nucleic Acids Research. 2000;28: 3442–3444. Available: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC110752/>
48. Aziz RK, Bartels D, Best AA, DeJongh M, Disz T, Edwards RA, et al. The RAST Server: Rapid Annotations using Subsystems Technology. BMC Genomics. 2008;9: 75. doi:10.1186/1471-2164-9-75
49. Overbeek R, Olson R, Pusch GD, Olsen GJ, Davis JJ, Disz T, et al. The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology

- (RAST). Nucleic Acids Research. 2014;42: D206–D214. doi:10.1093/nar/gkt1226
50. Medema MH, Fischbach MA. Computational approaches to natural product discovery. Nature Chemical Biology. 2015;11: 639–648. doi:10.1038/nchembio.1884
51. Noda-García L, Camacho-Zarco AR, Medina-Ruiz S, Gaytán P, Carrillo-Tripp M, Fülöp V, et al. Evolution of Substrate Specificity in a Recipient’s Enzyme Following Horizontal Gene Transfer. Molecular Biology and Evolution. 2013;30: 2024–2034. doi:10.1093/molbev/mst115
52. Carbonell P, Faulon J-L. Molecular signatures-based prediction of enzyme promiscuity. Bioinformatics. 2010;26: 2012–2019. doi:10.1093/bioinformatics/btq317
53. Cheng X-Y, Huang W-J, Hu S-C, Zhang H-L, Wang H, Zhang J-X, et al. A Global Characterization and Identification of Multifunctional Enzymes. PLoS ONE. 2012;7. doi:10.1371/journal.pone.0038979
54. Nagao C, Nagano N, Mizuguchi K. Prediction of Detailed Enzyme Functions and Identification of Specificity Determining Residues by Random Forests. PLOS ONE. 2014;9: e84623. doi:10.1371/journal.pone.0084623
55. Noda-García L, Juárez-Vázquez AL, Ávila-Arcos MC, Verduzco-Castro EA, Montero-Morán G, Gaytán P, et al. Insights into the evolution of enzyme substrate promiscuity after the discovery of  $\beta\alpha_8$  isomerase evolutionary intermediates from a diverse metagenome. BMC Evolutionary Biology. 2015;15. doi:10.1186/s12862-015-0378-1
56. Garcia-Seisdedos H, Ibarra-Molero B, Sanchez-Ruiz JM. Probing the Mutational Interplay between Primary and Promiscuous Protein Functions: A Computational-Experimental Approach. PLOS Computational Biology. 2012;8: e1002558. doi:10.1371/journal.pcbi.1002558
57. Nesvizhskii AI, Vitek O, Aebersold R. Analysis and validation of proteomic data generated by tandem mass spectrometry. Nature Methods. 2007;4: 787–797. doi:10.1038/nmeth1088
58. Campbell I. Biophysical Techniques - Paperback - Iain D. Campbell - Oxford University Press [Internet]. 2012. Available: <https://global.oup.com/ushe/product/biophysical-techniques-9780199642144?cc=mx&lang=en&>
59. Yang JY, Sanchez LM, Rath CM, Liu X, Boudreau PD, Bruns N, et al. Molecular Networking as a Dereplication Strategy. Journal of Natural Products. 2013;76: 1686–1699. doi:10.1021/np400413s
60. Köcher T, Superti-Furga G. Mass spectrometry-based functional proteomics: From molecular machines to protein networks. Nature Methods. 2007;4: 807–815.

- doi:10.1038/nmeth1093
61. James LC, Tawfik DS. Conformational diversity and protein evolution – a 60-year-old hypothesis revisited. *Trends in Biochemical Sciences*. 2003;28: 361–368. doi:10.1016/S0968-0004(03)00135-X
  62. Parisi G, Zea DJ, Monzon AM, Marino-Buslje C. Conformational diversity and the emergence of sequence signatures during evolution. *Current Opinion in Structural Biology*. 2015;32: 58–65. doi:10.1016/j.sbi.2015.02.005
  63. Javier Zea D, Miguel Monzon A, Fornasari MS, Marino-Buslje C, Parisi G. Protein Conformational Diversity Correlates with Evolutionary Rate. *Molecular Biology and Evolution*. 2013;30: 1500–1503. doi:10.1093/molbev/mst065
  64. Gatti-Lafranconi P, Hollfelder F. Flexibility and Reactivity in Promiscuous Enzymes. *ChemBioChem*. 2013;14: 285–292. doi:10.1002/cbic.201200628
  65. Cruz-Morales P, Kopp JF, Martínez-Guerrero C, Yáñez-Guerra LA, Selem-Mojica N, Ramos-Aboites H, et al. Phylogenomic Analysis of Natural Products Biosynthetic Gene Clusters Allows Discovery of Arseno-Organic Metabolites in Model Streptomyces. *Genome Biology and Evolution*. 2016;8: 1906–1916. doi:10.1093/gbe/evw125
  66. Li L, Stoeckert CJ, Roos DS. OrthoMCL: Identification of Ortholog Groups for Eukaryotic Genomes. *Genome Research*. 2003;13: 2178–2189. doi:10.1101/gr.1224503
  67. Waterhouse RM, Tegenfeldt F, Li J, Zdobnov EM, Kriventseva EV. OrthoDB: A hierarchical catalog of animal, fungal and bacterial orthologs. *Nucleic Acids Research*. 2013;41: D358–D365. doi:10.1093/nar/gks1116
  68. Gao B, Gupta RS. Phylogenetic Framework and Molecular Signatures for the Main Clades of the Phylum Actinobacteria. *Microbiology and Molecular Biology Reviews : MMBR*. 2012;76: 66–112. doi:10.1128/MMBR.05011-11
  69. Sen A, Daubin V, Abrouk D, Gifford I, Berry AM, Normand P. Phylogeny of the class Actinobacteria revisited in the light of complete genomes. The orders “Frankiales” and Micrococcales should be split into coherent entities: Proposal of Frankiales ord. nov., Geodermatophilales ord. nov., Acidothermales ord. nov. and Nakamurellales ord. nov. *International Journal of Systematic and Evolutionary Microbiology*. 2014;64: 3821–3832. doi:10.1099/ijjs.0.063966-0
  70. Zhou Z, Gu J, Li Y-Q, Wang Y. Genome plasticity and systems evolution in Streptomyces. *BMC Bioinformatics*. 2012;13: S8. doi:10.1186/1471-2105-13-S10-S8
  71. Kim J-N, Kim Y, Jeong Y, Roe J-H, Kim B-G, Cho B-K. Comparative Genomics Reveals the Core and Accessory Genomes of Streptomyces Species. *Journal of Microbiology and Biotechnology*. 2015;25: 1599–1605. doi:10.4014/jmb.1504.04008
  72. Nam H, Lewis NE, Lerman JA, Lee D-H, Chang RL, Kim D, et al. Network

- Context and Selection in the Evolution to Enzyme Specificity. *Science*. 2012;337: 1101–1104. doi:10.1126/science.1216861
73. Copley SD. An Evolutionary Biochemist’s Perspective on Promiscuity. *Trends in biochemical sciences*. 2015;40: 72–78. doi:10.1016/j.tibs.2014.12.004
74. Divergent Evolution of Enzymatic Function: Mechanistically Diverse Superfamilies and Functionally Distinct Suprafamilies. *Annual Review of Biochemistry*. 2001;70: 209–246. doi:10.1146/annurev.biochem.70.1.209
75. Huang R, Hippauf F, Rohrbeck D, Haustein M, Wenke K, Feike J, et al. Enzyme functional evolution through improved catalysis of ancestrally nonpreferred substrates. *Proceedings of the National Academy of Sciences*. 2012;109: 2966–2971. doi:10.1073/pnas.1019605109
76. Fondi M, Emiliani G, Liò P, Gribaldo S, Fani R. The evolution of histidine biosynthesis in archaea: Insights into the his genes structure and organization in LUCA. *Journal of Molecular Evolution*. 2009;69: 512–526. doi:10.1007/s00239-009-9286-6
77. Merino E, Jensen RA, Yanofsky C. Evolution of bacterial trp operons and their regulation. *Current opinion in microbiology*. 2008;11: 78–86. doi:10.1016/j.mib.2008.02.005
78. Verduzco-Castro EA, Michalska K, Endres M, Juárez-Vazquez AL, Noda-García L, Chang C, et al. Co-occurrence of analogous enzymes determines evolution of a novel  $\beta\alpha_8$ -isomerase sub-family after non-conserved mutations in flexible loop. *Biochemical Journal*. 2016;473: 1141–1152. doi:10.1042/BJ20151271
79. Noda-Garcia L. Estudio de la evolución molecular de la función enzimática susando como modelo una enzima con características ancestrales. PhD thesis, Langebio, CINVESTAV. 2012.
80. Petrenko R, Meller J. Molecular Dynamics. eLS. John Wiley & Sons, Ltd; 2001. Available: <http://onlinelibrary.wiley.com/doi/10.1002/9780470015902.a0003048.pub2/abstract>
81. Molecular Modeling of Proteins Andreas Kukol Springer [Internet]. Available: <http://www.springer.com/us/book/9781588298645>
82. Sikosek T, Chan HS. Biophysics of protein evolution and evolutionary protein biophysics. *Journal of The Royal Society Interface*. 2014;11: 20140419. doi:10.1098/rsif.2014.0419
83. Zhou R. Replica Exchange Molecular Dynamics Method for Protein Folding Simulation. In: Bai Y, Nussinov R, editors. *Protein Folding Protocols*. Humana Press; 2006. pp. 205–223. Available: <http://dx.doi.org/10.1385/1-59745-189-4%3A205>
84. Bisswanger H. General Aspects of Enzyme Analysis. *Practical Enzymology*.

- Wiley-VCH Verlag GmbH & Co. KGaA; 2011. pp. 5–91. Available: <http://onlinelibrary.wiley.com/doi/10.1002/9783527659227.ch2/summary>
85. Hommel U, Eberhard M, Kirschner K. Phosphoribosyl Anthranilate Isomerase Catalyzes a Reversible Amadori Reaction. *Biochemistry*. 1995;34: 5429–5439. doi:10.1021/bi00016a014
86. Scheer M, Grote A, Chang A, Schomburg I, Munaretto C, Rother M, et al. BRENDA, the enzyme information system in 2011. *Nucleic Acids Research*. 2011;39: D670–D676. doi:10.1093/nar/gkq1089
87. Van Der Spoel D, Lindahl E, Hess B, Groenhof G, Mark AE, Berendsen HJC. GROMACS: Fast, flexible, and free. *Journal of Computational Chemistry*. 2005;26: 1701–1718. doi:10.1002/jcc.20291
88. Odokonyero D, Sakai A, Patskovsky Y, Malashkevich VN, Fedorov AA, Bonanno JB, et al. Loss of quaternary structure is associated with rapid sequence divergence in the OSBS family. *Proceedings of the National Academy of Sciences of the United States of America*. 2014;111: 8535–8540. doi:10.1073/pnas.1318703111
89. Osbourn A. Gene Clusters for Secondary Metabolic Pathways: An Emerging Theme in Plant Biology. *Plant Physiology*. 2010;154: 531–535. doi:10.1104/pp.110.161315
90. Makarova KS, Aravind L, Galperin MY, Grishin NV, Tatusov RL, Wolf YI, et al. Comparative Genomics of the Archaea (Euryarchaeota): Evolution of Conserved Protein Families, the Stable Core, and the Variable Shell. *Genome Research*. 1999;9: 608–628. doi:10.1101/gr.9.7.608
91. Harvey AL, Edrada-Ebel R, Quinn RJ. The re-emergence of natural products for drug discovery in the genomics era. *Nature Reviews Drug Discovery*. 2015;14: 111–129. doi:10.1038/nrd4510
92. Benedict MN, Gonnerman MC, Metcalf WW, Price ND. Genome-Scale Metabolic Reconstruction and Hypothesis Testing in the Methanogenic Archaeon Methanosaerica acetivorans C2A. *Journal of Bacteriology*. 2012;194: 855–865. doi:10.1128/JB.06040-11
93. Seitz KW, Lazar CS, Hinrichs K-U, Teske AP, Baker BJ. Genomic reconstruction of a novel, deeply branched sediment archaeal phylum with pathways for acetogenesis and sulfur reduction. *The ISME Journal*. 2016;10: 1696–1705. doi:10.1038/ismej.2015.233
94. Moustafa A, Loram JE, Hackett JD, Anderson DM, Plumley FG, Bhattacharya D. Origin of Saxitoxin Biosynthetic Genes in Cyanobacteria. *PLOS ONE*. 2009;4: e5758. doi:10.1371/journal.pone.0005758
95. Medema MH, Osbourn A. Computational genomic identification and functional reconstitution of plant natural product biosynthetic pathways. *Natural Product*

- Reports. 2016;33: 951–962. doi:10.1039/c6np00035e
96. Medema MH, Kottmann R, Yilmaz P, Cummings M, Biggins JB, Blin K, et al. Minimum Information about a Biosynthetic Gene cluster. *Nature Chemical Biology*. 2015;11: 625–631. doi:10.1038/nchembio.1890
97. Iqbal HA, Low-Beinart L, Obiajulu JU, Brady SF. Natural Product Discovery through Improved Functional Metagenomics in Streptomyces. *Journal of the American Chemical Society*. 2016;138: 9341–9344. doi:10.1021/jacs.6b02921
98. Ulas T, Riemer SA, Zaparty M, Siebers B, Schomburg D. Genome-Scale Reconstruction and Analysis of the Metabolic Network in the Hyperthermophilic Archaeon *Sulfolobus Solfataricus*. *PLoS ONE*. 2012;7. doi:10.1371/journal.pone.0043401
99. Charlesworth JC, Burns BP. Untapped Resources: Biotechnological Potential of Peptides and Secondary Metabolites in Archaea. *Archaea*. 2015;2015: e282035. doi:10.1155/2015/282035
100. Computational Pan-Genomics Consortium. Computational pan-genomics: Status, promises and challenges. *Briefings in Bioinformatics*. 2016; doi:10.1093/bib/bbw089
101. Chan C, Jayasekera S, Kao B, Páramo M, Grotthuss M von, Ranz JM. Remodelling of a homeobox gene cluster by multiple independent gene reunions in *Drosophila*. *Nature Communications*. 2015;6: 6509. doi:10.1038/ncomms7509
102. Rinke C, Schwientek P, Sczyrba A, Ivanova NN, Anderson IJ, Cheng J-F, et al. Insights into the phylogeny and coding potential of microbial dark matter. *Nature*. 2013;499: 431–437. doi:10.1038/nature12352
103. Castelle CJ, Wrighton KC, Thomas BC, Hug LA, Brown CT, Wilkins MJ, et al. Genomic Expansion of Domain Archaea Highlights Roles for Organisms from New Phyla in Anaerobic Carbon Cycling. *Current Biology*. 2015;25: 690–701. doi:10.1016/j.cub.2015.01.014
104. Spang A, Saw JH, Jørgensen SL, Zaremba-Niedzwiedzka K, Martijn J, Lind AE, et al. Complex archaea that bridge the gap between prokaryotes and eukaryotes. *Nature*. 2015;521: 173–179. doi:10.1038/nature14447
105. Koonin EV. Archaeal ancestors of eukaryotes: Not so elusive any more. *BMC Biology*. 2015;13. doi:10.1186/s12915-015-0194-5
106. Chaudhari NM, Gupta VK, Dutta C. BPGA- an ultra-fast pan-genome analysis pipeline. *Scientific Reports*. 2016;6: 24373. doi:10.1038/srep24373
107. Glass JI, Assad-Garcia N, Alperovich N, Yooseph S, Lewis MR, Maruf M, et al. Essential genes of a minimal bacterium. *Proceedings of the National Academy of Sciences of the United States of America*. 2006;103: 425–430.

- doi:10.1073/pnas.0510013103
108. Narechania A, Baker RH, Sit R, Kolokotronis S-O, DeSalle R, Planet PJ. Random Addition Concatenation Analysis: A Novel Approach to the Exploration of Phylogenomic Signal Reveals Strong Agreement between Core and Shell Genomic Partitions in the Cyanobacteria. *Genome Biology and Evolution*. 2012;4: 30–43. doi:10.1093/gbe/evr121
  109. Stamatakis A. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*. 2014;30: 1312–1313. doi:10.1093/bioinformatics/btu033
  110. Powerful tree graphics with ggplot2 [Internet]. Available: [http://joey711.github.io/phyloseq/plot\\_tree-examples.html](http://joey711.github.io/phyloseq/plot_tree-examples.html)
  111. Zacharia VM, Traxler MF. Exploring new horizons. *eLife*. 2017;6: e23624. doi:10.7554/eLife.23624
  112. Woese C. The universal ancestor. *Proceedings of the National Academy of Sciences*. 1998;95: 6854–6859. Available: <http://www.pnas.org/content/95/12/6854>
  113. Woese CR, Gupta R. Are archaebacteria merely derived “prokaryotes”? *Nature*. 1981;289: 95–96. doi:10.1038/289095a0
  114. Woese CR, Kandler O, Wheelis ML. Towards a natural system of organisms: Proposal for the domains Archaea, Bacteria, and Eucarya. *Proceedings of the National Academy of Sciences of the United States of America*. 1990;87: 4576–4579.
  115. Woese CR, Fox GE. Phylogenetic structure of the prokaryotic domain: The primary kingdoms. *Proceedings of the National Academy of Sciences of the United States of America*. 1977;74: 5088–5090. Available: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC432104/>
  116. Woese CR. There must be a prokaryote somewhere: Microbiology’s search for itself. *Microbiological Reviews*. 1994;58: 1–9. Available: <http://mmbr.asm.org/content/58/1/1>
  117. Graham DE, Overbeek R, Olsen GJ, Woese CR. An archaeal genomic signature. *Proceedings of the National Academy of Sciences*. 2000;97: 3304–3308. doi:10.1073/pnas.97.7.3304
  118. Howland JL. The surprising archaea: Discovering another domain of life. New York: Oxford University; 2000.
  119. Xu Y, Gogarten JP. Computational Methods for Understanding Bacterial and Archaeal Genomes. World Scientific; 2008.
  120. Garrett RA, Klenk H-P. Archaea: Evolution, Physiology, and Molecular Biology.

- John Wiley & Sons; 2008.
121. Koonin EV, Mushegian AR, Galperin MY, Walker DR. Comparison of archaeal and bacterial genomes: Computer analysis of protein sequences predicts novel functions and suggests a chimeric origin for the archaea. *Molecular Microbiology*. 1997;25: 619–637. doi:10.1046/j.1365-2958.1997.4821861.x
  122. Koonin EV, Wolf YI. Genomics of bacteria and archaea: The emerging dynamic view of the prokaryotic world. *Nucleic Acids Research*. 2008;36: 6688–6719. doi:10.1093/nar/gkn668
  123. Koonin EV. The Turbulent Network Dynamics of Microbial Evolution and the Statistical Tree of Life. *Journal of Molecular Evolution*. 2015;80: 244–250. doi:10.1007/s00239-015-9679-7
  124. Land M, Hauser L, Jun S-R, Nookaew I, Leuze MR, Ahn T-H, et al. Insights from 20 years of bacterial genome sequencing. *Functional & Integrative Genomics*. 2015;15: 141–161. doi:10.1007/s10142-015-0433-4
  125. Nishida H. Evolution of genome base composition and genome size in bacteria. *Frontiers in Microbiology*. 2012;3. doi:10.3389/fmicb.2012.00420
  126. Coyle M, Hu J, Gartner Z. Mysteries in a Minimal Genome. *ACS Central Science*. 2016;2: 274–277. doi:10.1021/acscentsci.6b00110
  127. O'Meara B. CRAN Task View: Phylogenetics, Especially Comparative Methods. 2016; Available: <https://CRAN.R-project.org/view=Phylogenetics>
  128. Larsson J, Nylander JA, Bergman B. Genome fluctuations in cyanobacteria reflect evolutionary, developmental and adaptive traits. *BMC Evolutionary Biology*. 2011;11: 187. doi:10.1186/1471-2148-11-187
  129. Whitton BA. Ecology of Cyanobacteria II: Their Diversity in Space and Time. Springer Science & Business Media; 2012.
  130. Cohen GN. The biosynthesis of histidine and its regulation. *Microbial Biochemistry*. Springer Netherlands; 2004. pp. 225–230. Available: [http://link.springer.com/chapter/10.1007/978-1-4020-2237-1\\_29](http://link.springer.com/chapter/10.1007/978-1-4020-2237-1_29)
  131. Plach MG, Reisinger B, Sterner R, Merkl R. Long-Term Persistence of Bi-functionality Contributes to the Robustness of Microbial Life through Exaptation. *PLOS Genetics*. 2016;12: e1005836. doi:10.1371/journal.pgen.1005836
  132. Battistuzzi FU, Feijao A, Hedges SB. A genomic timescale of prokaryote evolution: Insights into the origin of methanogenesis, phototrophy, and the colonization of land. *BMC Evolutionary Biology*. 2004;4: 44. doi:10.1186/1471-2148-4-44
  133. Lapierre P, Gogarten JP. Estimating the size of the bacterial pan-genome.

- Trends in Genetics. 2009;25: 107–110. doi:10.1016/j.tig.2008.12.004
134. Větrovský T, Baldrian P. The Variability of the 16S rRNA Gene in Bacterial Genomes and Its Consequences for Bacterial Community Analyses. PLoS ONE. 2013;8. doi:10.1371/journal.pone.0057923
135. Brettin T, Davis JJ, Disz T, Edwards RA, Gerdes S, Olsen GJ, et al. RASTtk: A modular and extensible implementation of the RAST algorithm for building custom annotation pipelines and annotating batches of genomes. Scientific Reports. 2015;5: 8365. doi:10.1038/srep08365
136. chesterismay. Updated R Markdown thesis template [Internet]. Chester's R blog. 2016. Available: <https://chesterismay.wordpress.com/2016/09/01/updated-r-markdown-thesis-template/>
137. Barona-Gómez F, Cruz-Morales P, Noda-García L. What can genome-scale metabolic network reconstructions do for prokaryotic systematics? Antonie van Leeuwenhoek. 2012;101: 35–43. doi:10.1007/s10482-011-9655-1
138. Weber T, Blin K, Duddela S, Krug D, Kim HU, Bruccoleri R, et al. antiSMASH 3.0—a comprehensive resource for the genome mining of biosynthetic gene clusters. Nucleic Acids Research. 2015;43: W237–W243. doi:10.1093/nar/gkv437
139. Medema MH, Blin K, Cimermancic P, Jager V de, Zakrzewski P, Fischbach MA, et al. antiSMASH: Rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences. Nucleic Acids Research. 2011;39: W339–W346. doi:10.1093/nar/gkr466
140. Molina ST, Borkovec TD. The Penn State worry questionnaire: Psychometric properties and associated characteristics. In: Davey GCL, Tallis F, editors. Worrying: Perspectives on theory, assessment and treatment. New York: Wiley; 1994. pp. 265–283.
141. Reed College. LaTeX your document [Internet]. 2007. Available: <http://web.reed.edu/cis/help/LaTeX/index.html>
142. Noble SG. Turning images into simple line-art. Undergraduate thesis, Reed College. 2002.