

La ocurrencia y variación de la promiscuidad en rutas del metabolismo especializado son reveladas por herramientas bioinformáticas evolutivas.

---

Una tesis

presentada a

LANGEBIO-CINVESTAV

---

Para cumplir con los

requerimientos de la obtención del grado

de Doctor en Ciencias

---

Nelly Selém

Enero 2019



Aprobado por el laboratorio  
Evolución de la diversidad metabólica

---

Francisco Barona Gomez



# Agradecimientos

Quiero agradecer por todas las enseñanzas a mis compañeros de trabajo Hilda, Ernesto, Cristian, César, Karina Verdel, Lianet, Adriana Espinosa y Abraham Avelar. A manu y Vero por su interés en emprender proyectos. A Dany mi veranito , a Ernesto de keri, secretaria innovacion, DNABits.



# Prefacio

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



# Table of Contents

<b>Introducción</b>	1
0.0.1 Funcion biologica de la promiscuidad enzimatica	2
0.0.2 Relacion del pangenooma con la promiscuidad enzimatica	3
0.0.3 Modelos bioinformaticos de promiscuidad	3
0.0.4 Promiscuidad in vitro y promiscuidad in vivo	4
0.0.5 El papel de la dinamica molecular en la promiscuidad	5
0.0.6 Modelo biológico diversidad de Actinobacteria	5
0.0.7 Modelo metabolico biosintesis de aminoacidos.	6
0.0.8 Modelo biológico	7
0.0.9 Subsistemas metabolicos	7
0.1 Modelos computacionales	7
<b>Antecedentes</b>	9
0.2 La promiscuidad puede abordarse a distintos niveles incluyendo enzima, familia y ruta de metabólica.	10
0.3 Antecedentes conceptuales	10
0.4 El establecimiento de un marco de conservación permite distinguir cambios	12
0.5 La genómica comparativa como herramienta en la distinción de familias y enzimas promiscas que participan en el metabolismo especializado.	13
0.6 La genómica comparativa como herramienta en la priorización de clusters promiscuos	14
0.6.1 Expansion y contextos genomicos como herramienta de anotacion funcional	14
0.6.2 Contexto y vecindades genomicas	16
0.7 Estudio de una familia PriA	16
0.7.1 Caracterizacion in vivo	16
0.7.2 Caracterizacion bioquimica in vitro.	17
0.7.3 Modelado de dinamica molecular	17
<b>Pregunta biológica</b>	19
<b>Objetivos</b>	21
0.8 Objetivo General	21
0.9 Objetivos particulares	21

<b>Estrategias</b>	23
Obtener informacion genomica del phylum Actinobacteria.	23
0.9.1 Anotar consistentemente las secuencias codificantes de estos genomas.	23
Establecer las relaciones filogeneticas de los genomas colectados.	23
0.9.2 La promiscuidad en familias enzimaticas.	23
Promiscuidad in vitro dentro de miembros de una familia promiscua de enzimas.	24
Sistematizar Evomining para convertirla una plataforma descargable y utilizable en cualquier set de datos bacterianos relacionados taxonomicamente proporcionados por el usuario.	24
Seleccionar miembros homologos de la familia de enzimas.	24
Medir cineticas enzimaticas, contexto genomico, vecindad genomica, flexibilidad y numero de conformeros.	24
Determinar posibles correlaciones entre los datos producidos.	24
0.9.3 Desarrollar una metodologia para la deteccion in vivo de promiscuidad enzimatica.	25
<b>Metodología</b>	27
0.9.4 La promiscuidad en familias enzimaticas.	27
Actinobacteria genomica	27
Annotation	27
Genomic DB phylogeny	27
0.9.5 Identificar cambios en la vecindad genomica en familias selectas de enzimas de metabolismo central.	28
Organizar y presentar los datos en una plataforma.	29
0.9.6 Promiscuidad in vitro	30
Datos cineticos:	30
Dinamica molecular	30
0.9.7 Promiscuidad in vivo	31
0.9.8 Consideraciones	31
<b>Chapter 1: Desarrollo de Orthocore e implementación de otras herramientas computacionales para entender el pangenoma de un linaje genómico.</b>	35
1.1 La distribución de la función metabólica de las familias del pangenoma depende de la variabilidad del linaje seleccionado.	36
1.2 El core conservado permite la reconstrucción de filogenias complicadas	40
1.3 El algoritmo de Orthocore	40
1.3.1 Componentes técnicos de Orthocore	41
1.3.2 Aplicaciones de Orthocore, identificación del core conservado y de familias de genes marcadores.	42
1.3.3 Genes de islas de patogenicidad de <i>Salmonella</i> en México están conservados en la mayoría de los genomas.	43

1.3.4	Nostoc provenientes del metagenoma de cíadas se agrupan Cyanobacteria . . . . .	43
1.3.5	Identificación de genes marcadores de <i>Clavibacter michiganensis</i>	45
1.4	Clavisual: Identificación de genes marcadores a un cierto porcentaje de grupos seleccionados . . . . .	47
1.4.1	El pangenoma de <i>Clavibacter Michiganensis</i> es abierto . . . . .	49
1.5	Relación entre genes marcadores, Orthocores y la promiscuidad enzimática. . . . .	50
1.6	Protocolos para usar Orthocore, myRAST, fastOrtho, Clavigenomics, y BPGA . . . . .	50
1.7	ORthocore . . . . .	53
<b>Chapter 2: EvoMining</b>	. . . . .	<b>55</b>
<b>Chapter 3: EvoMining</b>	. . . . .	<b>63</b>
3.1	Introduction . . . . .	63
3.2	EvoMining es un método para encontrar BGCs no tradicionales . . . . .	71
3.3	Gen families expansions on genomes . . . . .	71
3.3.1	Pangenomes . . . . .	71
3.4	EvoMining . . . . .	71
3.5	Pangenome . . . . .	71
3.6	EvoMining Implementation . . . . .	72
3.7	EvoMining Databases . . . . .	73
Genome DB . . . . .	73	
Phylogeny . . . . .	74	
Central DB . . . . .	74	
3.8	Data Bases . . . . .	75
3.8.1	Central pathways . . . . .	75
3.8.2	Genome Dynamics . . . . .	75
Natural Products DB . . . . .	76	
3.8.3	AntisMASH optional DB . . . . .	77
3.8.4	Otras estrategias para los clusters Argon context Idea . . . . .	77
3.9	Argonne . . . . .	77
3.9.1	Inline code . . . . .	80
3.10	Recomendaciones de Luis . . . . .	80
3.11	CORASON: Other genome Mining tools context-based . . . . .	81
3.12	CORe Analysis of Syntenic Orthologs to prioritize Natural Product-Biosynthetic Gene Cluster . . . . .	81
3.13	Tree methods (from antiSMASH textual quotation) . . . . .	82
<b>Chapter 4: EvoMining Results</b>	. . . . .	<b>83</b>
4.1	Archaea . . . . .	83
4.2	Tables . . . . .	84
4.2.1	Expansions BoxPlot by metabolic family . . . . .	85
4.2.2	Expansions BoxPlot by metabolic family by phylum . . . . .	86

4.3	Central pathway expansions . . . . .	97
4.4	Genome Size correlations . . . . .	99
4.4.1	Correlation between genome size and AntiSMASH products .	99
4.4.2	Correlation between genome size and Central pathway expansions	101
4.5	Natural products . . . . .	105
4.5.1	Natural products recruitments from EvoMining heatplot . .	105
4.6	Archaeas AntiSMASH . . . . .	107
4.6.1	AntisSMASH vs Central Expansions . . . . .	109
4.7	Selected trees from EvoMining . . . . .	112
4.8	. . . . .	113
4.9	Bibliographies . . . . .	113
4.10	Anything else? . . . . .	114
4.11	Actinobacteria . . . . .	114
4.12	Tables . . . . .	114
4.12.1	Expansions BoxPlot by metabolic family . . . . .	116
4.13	Central pathway expansions . . . . .	118
4.14	Genome Size correlations . . . . .	122
4.14.1	Correlation between genome size and AntiSMASH products .	122
4.14.2	Correlation between genome size and Central pathway expansions	124
4.15	Natural products . . . . .	128
4.15.1	Natural products recruitments from EvoMining heatplot . .	128
4.16	Actinos AntiSMASH . . . . .	130
4.16.1	AntisSMASH vs Central Expansions . . . . .	132
4.17	Selected trees from EvoMining . . . . .	135
4.18	Cyanobacteria . . . . .	137
4.19	Tables . . . . .	137
4.19.1	Expansions BoxPlot by metabolic family . . . . .	138
4.20	Central pathway expansions . . . . .	140
4.21	Genome Size correlations . . . . .	141
4.21.1	Correlation between genome size and AntiSMASH products .	141
4.21.2	Correlation between genome size and Central pathway expansions	143
4.22	Natural products . . . . .	147
4.22.1	Natural products recruitments from EvoMining heatplot . .	147
4.23	Cyanobacterias AntiSMASH . . . . .	149
4.23.1	AntisSMASH vs Central Expansions . . . . .	151
4.24	Selected trees from EvoMining . . . . .	154
<b>Chapter 5: CORASON . . . . .</b>	<b>159</b>	
<b>Chapter 6: Desarrollo de CORASON como herramienta para organizar clusters biosintéticos y otras vecindades genómicas conservadas. .</b>	<b>161</b>	
6.1	Algoritmo y características de CORASON . . . . .	162
6.2	Las familias de BGCs están formadas por variantes del BGC de referencia.	162
6.3	Aplicaciones de CORASON en Actinobacteria y Pseudomonas . . . .	163
6.4	Estas familias pueden clasificarse . . . . .	164

6.5	Un BGC reportado puede tener muchas variantes que conforman una familia de BGCs . . . . .	164
6.6	CORASON analiza la conservación del contexto genómico de los Evo-Mining hits . . . . .	166
6.7	El contexto de una rama divergente de <i>tauD</i> está conservado en Pseudomonas . . . . .	166
6.8	CORASON se adaptó a la herramienta BiG-SCAPE de clasificación de familias de BGCs . . . . .	166
6.9	BiG SCAPE y CORASON identificaron nuevos productos variantes de la familia de BGCs Rimosamide - Detoxin en Actinobacteria . . . . .	167
<b>Chapter 7: CORASON sugiere que las familias detoxin y rimosamide pertenecen a un amplio clan de familias dedicadas a la síntesis de péptidos.</b>	. . . . .	<b>169</b>
7.1	Clados selectos del árbol de CORASON que contiene a las familias rimosamide/detoxin contienen diversidad génica que correlaciona con novedad química. . . . .	170
7.1.1	El superclado spectinomycin/ detoxin-rimosamide-clan produce cinco variantes de detoxin. . . . .	170
7.1.2	El clado <i>Amycolatopsis P450</i> produce cinco variantes de detoxin.	171
<b>Chapter 8: Discusión</b>	. . . . .	<b>173</b>
<b>Conclusion</b>	. . . . .	<b>177</b>
4.1	Discussion . . . . .	177
More info	. . . . .	177
<b>Appendix A: The First Appendix</b>	. . . . .	<b>179</b>
In the main Rmd file:	. . . . .	179
In :	. . . . .	179
<b>Appendix B: The Second Appendix, Open source code on this document</b>	. . . . .	<b>181</b>
B.1	R markdown . . . . .	181
B.2	Docker . . . . .	181
B.3	Git . . . . .	182
B.4	Connect GitHub and DockerHub . . . . .	182
B.5	Additional resources . . . . .	182
<b>Appendix C: The third Appendix, Other contributions during my phd</b>	. . . . .	<b>185</b>
C.1	Accepted . . . . .	185
C.2	Submitted . . . . .	185
C.3	On preparation . . . . .	185
<b>References</b>	. . . . .	<b>187</b>



# List of Tables

1.1	Correlation of Inheritance Factors for Parents and Child . . . . .	52
3.1	BBH_Organisms . . . . .	75
3.2	WC_Organisms . . . . .	75
4.1	Families on Archaeabacteria . . . . .	84
4.2	Correlation of Inheritance Factors for Parents and Child . . . . .	114
4.3	Families on Cyanobacteria . . . . .	137



# List of Figures

1	Antecedentes Conceptuales . . . . .	9
2	Antecedentes Conceptuales . . . . .	11
1.1	Orthocore calcula el core de un linaje genómico para proveer una filogenia	35
1.2	El metabolismo en el Pangenoma . . . . .	37
1.3	Core y genes Marcadores . . . . .	39
1.4	Los mejores hits n-direccionales generalizan a los <i>Bidirectional Best Hits</i>	41
1.5	Arbol filogenético de <i>Nostoc</i> construido utilizando la selección de genes del <i>core conservado</i> . . . . .	44
1.6	Los genes del cluster biosintético clv son marcadores de <i>Clavibacter</i> .	45
1.7	clv BGC . . . . .	46
1.8	X X Evomining . . . . .	48
1.9	X X XXXX . . . . .	49
1.10	X X . . . . .	50
1.11	X X . . . . .	51
3.1	EvoMining Algorithm . . . . .	64
3.2	Seed genomes . . . . .	65
3.3	Expansion patterns in 42 conserved families . . . . .	66
3.4	EvoMining Algorithm . . . . .	67
3.5	EvoMining Algorithm . . . . .	68
3.6	EvoMining Algorithm . . . . .	69
3.7	EvoMining Algorithm . . . . .	70
4.1	Expansions Boxplot . . . . .	85
4.2	Archaeas Heatplot . . . . .	98
4.3	Correlation between Archaeas genome size and antimash Natural products detection colored by Order . . . . .	99
4.4	Correlation between Archaeas genome size and central pathway expansions . . . . .	101
4.5	Correlation between Archaeas genome size and central pathway expansions grided by order . . . . .	102
4.6	Correlation between Archaeas Genome size vs Total central pathway expansion coloured by metabolic Family . . . . .	103
4.7	Archaeas Recruitmens on central families coloured by kingdom . . . .	105
4.8	Archaeas Recruitmens on central families coloured by taxonomy . . .	106

4.9	Archaeas Diversity . . . . .	107
4.10	Archaeas Smash Taxonomical Diversity . . . . .	108
4.11	Correlation between Archaeas central pathway expansions and anti-smash Natural products detection . . . . .	109
4.12	Correlation between Archaeas central pathway expnasions and anti-smash Natural products detection . . . . .	110
4.13	Archaeas Natural products by family . . . . .	111
4.14	Phosphoribosyl isomerase A EvoMiningtree . . . . .	112
4.15	Phosphoribosyl isomerase other EvoMiningtree . . . . .	112
4.16	Phosphoribosyl anthranilate isomerase EvoMiningtree . . . . .	112
4.17	Expansions Boxplot . . . . .	117
4.18	Actinobacterial Heatplot . . . . .	119
4.19	Streptomyces Genomes expansions on PGA Aminoacids HeatPlot . . . . .	121
4.20	Correlation between Actinos genome size and antismash Natural products detection colored by Order . . . . .	122
4.21	Correlation between Actinos genome size and antismash Natural products detection grided by Order . . . . .	123
4.22	Correlation between Actinos genome size and central pathway expansions . . . . .	124
4.23	Correlation between Actinos genome size and central pathway expansions grided by order . . . . .	125
4.24	Correlation between Actinos Genome size vs Total central pathway expansion coloured by metabolic Family . . . . .	126
4.25	Actinos Recruitmens on central families coloured by kingdom . . . . .	128
4.26	Actinos Recruitmens on central families coloured by taxonomy . . . . .	129
4.27	Actinos Diversity . . . . .	130
4.28	Actinos Smash Taxonomical Diversity . . . . .	131
4.29	Correlation between Actinos central pathway expansions and antismash Natural products detection . . . . .	132
4.30	Correlation between Actinos central pathway expnasions and antismash Natural products detection . . . . .	133
4.31	Actinos Natural products by family . . . . .	134
4.32	Enolase EvoMiningtree . . . . .	135
4.33	Phosphoribosyl isomerase EvoMiningtree . . . . .	135
4.34	Phosphoribosyl isomerase A EvoMiningtree . . . . .	135
4.35	phosphoshikimate carboxyvinyltransferase EvoMiningtree . . . . .	136
4.36	Expansions Boxplot . . . . .	139
4.37	Cyanobacterial Heatplot . . . . .	140
4.38	Correlation between genome size and antismash Natural products detection colored by Order . . . . .	141
4.39	Correlation between genome size and antismash Natural products detection grided by Order . . . . .	142
4.40	Correlation between genome size and central pathway expansions . . . . .	143
4.41	Correlation between genome size and central pathway expansions grided by order . . . . .	144

4.42 Correlation between Genome size vs Total central pathway expansion coloured by metabolic Family . . . . .	145
4.43 Recruitmens on central families coloured by kingdom . . . . .	147
4.44 Recruitmens on central families coloured by taxonomy . . . . .	148
4.45 Diversity . . . . .	149
4.46 Smash . . . . .	150
4.47 Correlation between central pathway axpnasions and antismash Natural products detection . . . . .	151
4.48 Correlation between central pathway axpnasions and antismash Natural products detection . . . . .	152
4.49 Natural products by family . . . . .	153
4.50 Phosphoribosyl isomerase EvoMiningtree . . . . .	154
4.51 Phosphoglycerate dehydrogenase EvoMiningtree . . . . .	154
4.52 Phosphoserine aminotransferase EvoMiningtree . . . . .	154
4.53 Triosephosphate isomerase EvoMiningtree . . . . .	155
4.54 glyceraldehyde3phosphate dehydrogenase EvoMiningtree . . . . .	155
4.55 phosphoglycerate kinase EvoMiningtree . . . . .	155
4.56 phosphoglycerate mutaseEvoMiningtree . . . . .	156
4.57 enolase EvoMiningtree . . . . .	156
4.58 Pyruvate kinase EvoMiningtree . . . . .	156
4.59 Aspartate transaminase EvoMiningtree . . . . .	157
4.60 Asparagine synthase EvoMiningtree . . . . .	157
4.61 Aspartate kinase EvoMiningtree . . . . .	157
4.62 Aspartate semialdehyde dehydrogenase EvoMiningtree . . . . .	158
4.63 Homoserine dehydrogenase EvoMiningtree . . . . .	158
8.1 EvoMining Algorithm . . . . .	174
8.2 EvoMining Algorithm . . . . .	175
8.3 EvoMining Algorithm . . . . .	176



# **Abstract**

The preface pretty much says it all.

Second paragraph of abstract starts here.



# **Dedicatoria**

You can have a dedication here if you wish.



# Introducción

Las enzimas catalizan reacciones químicas transformando sustratos en productos. During 20th century enzymes were percived as highly specific catalizer, 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 es necesario contar con una definición, algunos autores emplean el término promiscuidad para describir actividades enzimáticas distintas a la función principal [1] otros lo ven como una actividad secundaria fortuita [2] que pudo aparecer de forma accidental o inducida artificialmente [3]. Otros más, cuando una enzima puede operar sobre un amplio rango de sustratos, prefieren llamarla multiespecífica [1]. A la acción de realizar distintas funciones catalíticas, ya sea al catalizar varias reacciones químicas o bien una misma reacción en sustratos diferentes se le conoce como promiscuidad enzimática [4]. Existen varios tipos de promiscuidad enzimática.

Por sustrato cuando la reacción es la misma pero se lleva a cabo en distintos sustratos ejemplo la familia PriA [5] y la familia de betalactamasas [6]

Catalítica cuando la enzima utiliza diferentes mecanismos de reacción y/o residuos catalíticos, e.g. la quimotripsina puede catalizar reacciones de amidasa y fosfotriesterasa en un mismo sitio activo. [4] Por condiciones del entorno, cuando la enzima cambia su conformación dependiendo de las condiciones químicas y físicas presentes como pH, temperatura, solventes orgánicos y salinidad e.g. algunas lipasas pueden actuar como sintetizadoras de ésteres en lugar de hidrolasas en presencia de solventes orgánicos [3].

Este trabajo se enfocará a la promiscuidad por sustrato, entendiendo así que la enzima es capaz de catalizar la misma reacción química en al menos dos sustratos. La promiscuidad por sustrato es importante en términos 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 [8]. 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 [9]

Si bien existen familias de enzimas con alta especificidad por sustrato, otras familias como el citocromo P450 [11] y las beta lactamasas [13] 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 [14]. Esta vision ha cambiado debido al conocimiento de mas enzimas con multifuncionalidad [15], sin afectar la eficiencia catalitica por la funcion primaria [16]. En 1976 el interes por la promiscuidad comenzó por su influencia en la evolucion de la funcion enzimatica[17], las aproximaciones variaron desde la aparicion de la sintesis funcional [19], 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 [20]. Entre 2005 y 2010 se avanza del estudio de una sola familia enzimatica 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 prediccion 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.

### 0.0.1 Funcion biologica de la promiscuidad enzimatica

¿Por que existe la promiscuidad enzimatica? 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 [16] mediante la adquisicion de nuevas funciones quimicas. Respecto a la robustez, se probó que sobreexpresar enzimas promiscuas puede rescatar perdidas genicas [27]. 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 [5]. 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[1].

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 [28] y pueden servir como bloques de construccion para vias metabolicas nuevas [31]. La respuesta inmediata de adaptacion de un organismo podria ser una consecuencia de su grado de promiscuidad.

### **0.0.2 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 [34]. 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 [14] 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 [31]. Para evitar problemas tecnicos del calculo del pangenoma existen otros modelos de medicion de variabilidad del genomica entre especies bacterianas [35].

### **0.0.3 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 [36]. 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 metodos de random forest y soporte vectorial en 6799 proteinas provenientes de la base de datos Universal Protein Resource (UniProt). Las enzimas son descritas con subsecuencias de aminoacidos incorporando ademas características biofisicas 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 metodos es que hay familias de enzimas con alta identidad de secuencia entre sus miembros, con cambios bruscos en promiscuidad, debidos por ejemplo a la dinamica genomica [41], lo que dificulta que considerar solo la secuencia conlleve 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 ademas 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 mecanisticos de las enzimas.

Otra limitante a los enfoques descritos es que mezclan en su conjunto de entrenamiento fenomenos distintos de promiscuidad. Cheng p. g. incluye enzimas moonlight que si bien poseen funciones adicionales a la catalización, son distantes a las enzimas promiscuas [2]. Ademas en ambos casos mezclan en el mismo conjunto enzimas bacterianas y eucariotas, con lo que si existia una huella basada en secuencia entonces esta puede diluirse por la gran distancia taxonomica entre estos grupos (Tabla 1).

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

La ganancia de promiscuidad no solo puede entenderse como la capacidad de convertir mas sustratos [36], sino tambien como la mejora de la capacidad catalitica respecto a ellos. El I-index [12], esta definido como un rango de valores entre 0 y 1 que tiende a 1 entre mas parecida sea la actividad de la enzima sobre distintos sustratos, la capacidad catalitica es medida en terminos del cociente de Michaelis - Menten  $\frac{K_{cat}}{K_m}$ . El indice ha sido utilizado para predecir la afinidad por sustrato del citocromo P450 [22]. Una limitante del indice  $I$  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  $K_{cat}, K_m$  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 analisis metabolicos [43] como los empleados en la detección de rutas no conservadas en la biosíntesis de productos naturales [44]. En particular para este fin se ha utilizado

espectrometria de masas MS/MS, [43] combinada con molecular networking para identificar productos similares [46]

### **0.0.5 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. [48]. 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 [49]. Se postula que un aspecto de la dinamica molecular relevante para la diversificacion de especificidad por sustrato es el numero de conformeros [50]. 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* [41].

Las regiones rigidas 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 [2]. 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 [51]. 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 [13].

### **0.0.6 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 [52]. 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 [53], en particular en Actinobacteria para identificar

relaciones entre las familias del phylum, se obtuvieron arboles multilocus de entre 100 y 157 genomas [55]. Estos estudios sugieren como separar los genomas disponibles para hacer el calculo de grupos de ortología. Finalmente, se han realizado estudios de plasticidad genómica en Streptomyces considerando 5 y 17 organismos de los 300 genomas disponibles en la actualidad [57] donde reportan 2,018 familias en el core genome y 32,574 en el pangenoma.

### 0.0.7 Modelo metabólico biosíntesis de aminoácidos.

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 metabólicos como metabolismo de carbohidratos o de lipidos; de estos subsistemas uno de los mas amplios es el metabolismo de aminoácidos con entre 429 y 910 secuencias segun el organismo. La síntesis de aminoácidos 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 función de HisA como la de TrpF [5]. La cantidad de familias en el subsistema de metabolismo de aminoácidos, su variabilidad, su conservación entre distintos grupos taxonómicos y la existencia de estos ejemplos en Actinobacteria lo posicionan como un buen punto de partida para la búsqueda de promiscuidad tanto de familias promiscuas como de miembros promiscuos de las mismas.

En las cuatro décadas de estudio de la promiscuidad enzimática, hemos aprendido que es un fenómeno distribuido en distintos subsistemas metabólicos [59] y que su existencia puede deberse tanto al desarrollo de nuevas funciones para fines adaptativos [17], como al rescate de una función perdida [27]. Por ello la dinámica de pérdida y ganancia de genes asociada al contexto genómico en bacterias se relaciona con cambio en la función enzimática [61]. Precisando, respecto a la ganancia de genes, se postula que la bifuncionalidad precede la duplicación [62]. Lo que implica que dada una duplicación muy posiblemente previamente la promiscuidad estuvo presente [63].

Se han desarrollado técnicas bioquímicas y metabólicas de medición [12], así como algoritmos computacionales de predicción de promiscuidad [36]. Un aspecto a mejorar dentro del modelado es la restricción del conjunto de estudio a un grupo taxonómico tan reducido que exista congruencia en las familias de ortología 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 [42], es posible que dentro de las técnicas computacionales la flexibilidad durante la dinámica molecular esté correlacionada con la promiscuidad de los miembros de una familia [48].

### 0.0.8 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).

### 0.0.9 Subsistemas metabólicos

Los operones his y trp de histidina [65] y triptofano [66] 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 [67]. En otras Actinobacterias trpF se pierde totalmente y la familia homologa de HisA, se vuelve promiscua [5] 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 [41] 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.

## 0.1 Modelos computacionales

Para desarrollar herramientas se adoptó el enfoque de los contenedores bioinformáticos, todo el código fue depositado y documentado en GitHub y distribuido a través de un contenedor Docker



# Antecedentes

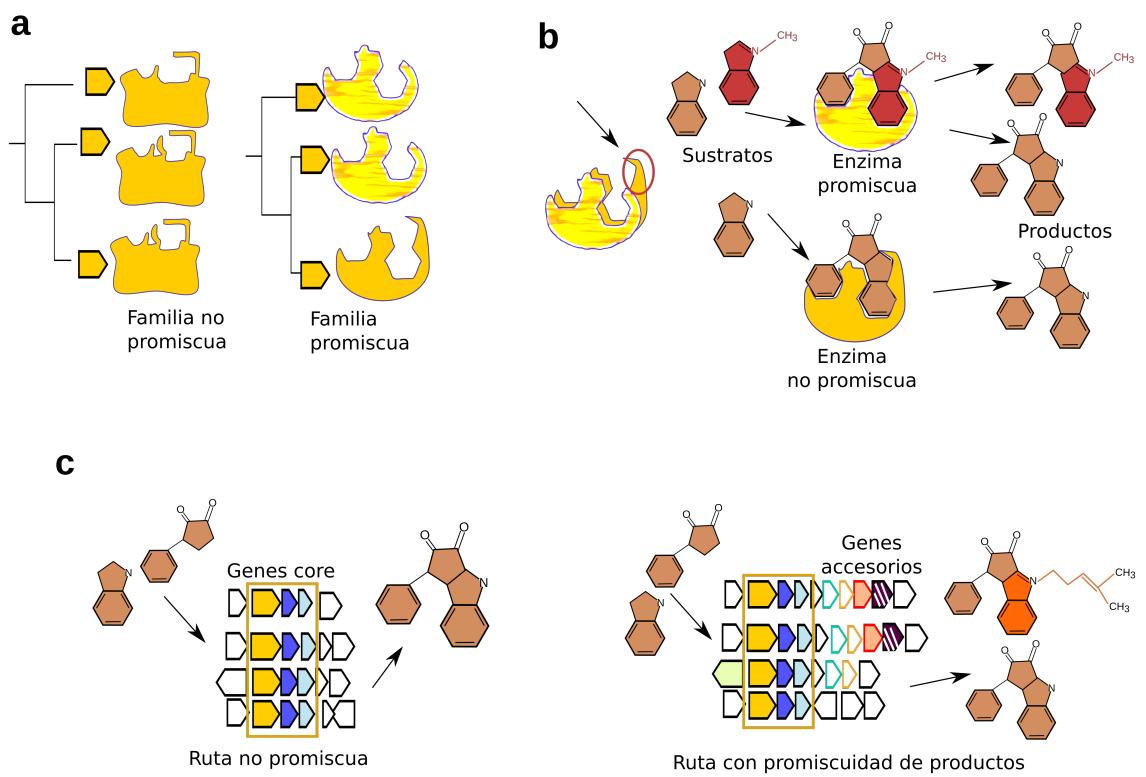


Figure 1: Antecedentes Conceptuales

## 0.2 La promiscuidad puede abordarse a distintos niveles incluyendo enzima, familia y ruta de metabólica.

Si se entiende a la promiscuidad como funciones alternativas de alguna unidad molecular, puede observarse promiscuidad a distintos niveles: desde un mismo gen que presenta splicing alternativo, una misma enzima con funciones alternativas, o bien una familia enzimática donde al menos algunos homólogos codifican enzimas promiscuas [20]. Existen también rutas que generan productos metabólicos alternativos[68], si se considera como la unidad de estudio a una ruta biosintética podemos generalizar la noción de promiscuidad al concepto de rutas promiscuas. La *Figura 1* muestra tres niveles en los que se puede estudiar la promiscuidad: i) Distinguendo familias de enzimas promiscuas de familias especialistas, ii) Distinguendo enzimas específicas de enzimas especialistas en una familia enzimática promiscua y finalmente iii) encontrando promiscuidad en rutas de metabolismo especializado.

PriA en Actinobacteria y HisA en enterobacteria son familias que ambas isomerizan proFAR en la ruta de síntesis de histidina pero solo la familia PriA es promiscua pues puede además isomerizar el sustrato PRA durante la síntesis de triptofano [5]. Sin embargo dentro de Actinobacteria, existen miembros no promiscuos de PriA, en algunas especies de Actinomyces los homólogos de *priA* codifican para enzimas monofuncionales en alguno de los dos sustratos, al menos en análisis *in vitro* [70]. Así pues dentro de una familia promiscua no todos los miembros tienen esta propiedad.

En este trabajo se buscará encontrar marcas de promiscuidad a nivel familia, enzima y ruta biosintética en el metabolismo especializado.

## 0.3 Antecedentes conceptuales

Se ha intentado identificar enzimas promiscuas mediante aprendizaje máquina utilizando únicamente la secuencia de aminoácidos. Estos enfoques no han distinguido entre identificación de familias promiscuas e identificación a nivel de enzima. [36] Hasta ahora al utilizar únicamente la información de la secuencia no ha sido posible identificar una familia promiscua sin conocer previamente al menos un miembro promiscuo de ella. Por otra parte, diferenciar la promiscuidad a nivel de enzima se dificulta cuando la identidad de secuencia es alta, como en el caso de las PriA que han perdido la promiscuidad en Actinomyces [70]

Para mejorar nuestro entendimiento del fenómeno además de la comparación de secuencias es necesario integrar otros elementos al análisis, Figura 2. Es difícil medir la promiscuidad en términos absolutos, por ejemplo, no se puede aseverar que una enzima es no promiscua sin haber previamente descartado todos los posibles sustratos del universo químico. Además incluso enzimas que resultan no promiscuas en análisis

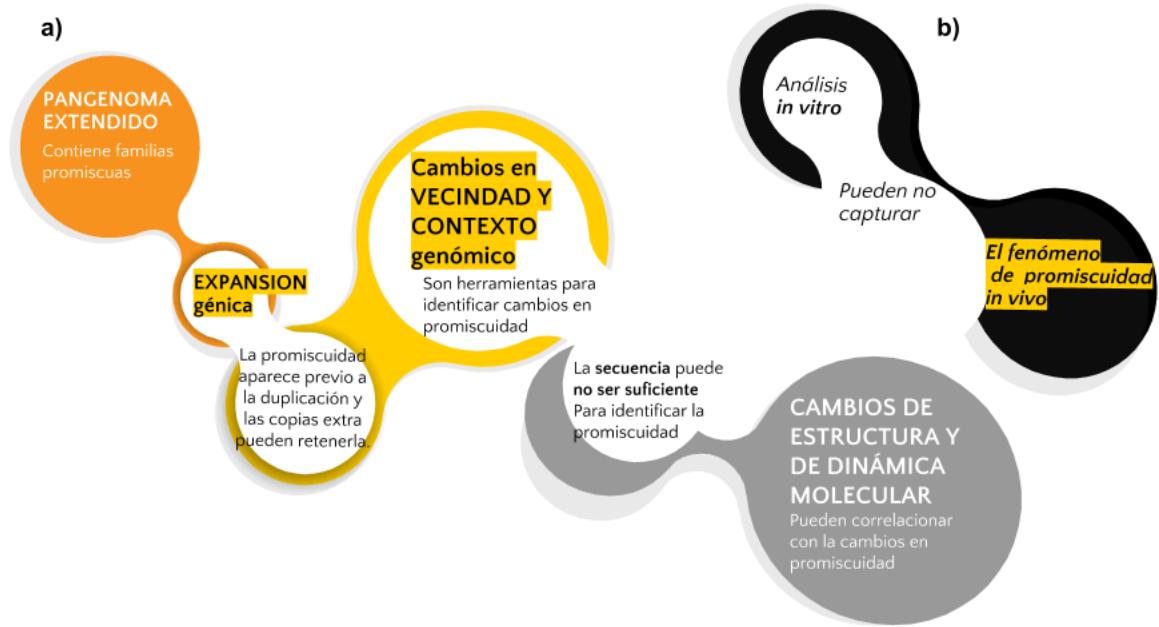


Figure 2: Antecedentes Conceptuales

in vivo si son promiscuas al examinarlas in vivo [71]. Sin embargo debido a que al adquirir una nueva función existe un umbral donde la función ancestral es conservada, es plausible estudiar transformar el problema de encontrar promiscuidad al de encontrar cambios en promiscuidad al relacionar estos últimos con las huellas que dejan los cambios funcionales[33]. Entre los elementos relevantes que correlacionan con la adquisición de una función alternativa se encuentran además de divergencia de secuencia, diversidad en vecindad genómica[72], la perdida o ganancia de genes[70], las expansiones genicas, i.e. el crecimiento del pangenoma endtro de un grupo taxonómico [26] y finalmente cambios estructurales o de flexibilidad durante la dinámica molecular[13]. Estos elementos tienen en común que reflejan un cambio en alguna propiedad genómica o biofísica observable en el registro evolutivo, de lo que se deriva que el buscar cambios en la promiscuidad de una enzima, familia o ruta, resulta mas factible por ahora que la búsqueda intrínseca de promiscuidad.

Debido a la abundancia de datos genómicos los primeros capítulos de este trabajo se centran en encontrar variaciones en secuencia, distribución del pangenoma, y vecindades genómicas para encontrar candidatos de familias, enzimas y rutas promiscuas.

## 0.4 El establecimiento de un marco de conservación permite distinguir cambios

La función de una enzima es un concepto jerárquico, dependiente de la filogenia de un organismo [73]. Por ello, para poder encontrar marcas de cambio funcional, por ejemplo diferencias en número de copias de una familia, primero es importante trabajar en la construcción de un marco filogenético consistente que permita ordenar inclusive organismos de la misma especie. La dificultad de esta tarea consiste en que si los organismos que se desea ordenar son muy cercanos, marcadores clásicos como el 16s son también muy parecidos en secuencia y no permiten resolver las relaciones entre ellos. Este caso dificultó la construcción de un árbol filogenético de *Actinomyces* y con ello se imposibilitaba encontrar patrones en la matriz de presencia / ausencia de genes [70].

Para solventar la falta de resolución de genes individuales en organismos cercanos puede utilizarse el conjunto de todos los genes comunes en un linaje. Este conjunto es conocido como el core genome. Se han desarrollado herramientas bioinformáticas para este problema, por ejemplo phyloPhlan fija 400 genes comunes en bacteria y trata de localizarlos dado un conjunto de genomas sin importar su linaje [74]. Sin embargo el contenido de genomas procariontes suele ser muy variable debido a mecanismos como transferencia horizontal y duplicación génica [75]. Esta variabilidad puede dificultar encontrar muchos de estos 400 genes o bien puede ser que en cierto linaje no sean tan informativos. Entre organismos de la misma especie pueden suceder fenómenos como que el core siempre se reduzca al aumentar un nuevo genoma y que el conjunto total de familias genéticas (pangenoma) siempre aumente. Aunado a esta observación biológica están también las limitaciones técnicas, hay genes que no aparecen en un genoma porque este fue mal secuenciado o ensamblado y por lo tanto estos genes disminuyen el tamaño del core.

Así pues para distinguir cambios genómicos ayuda establecer primero un orden entre los genomas del linaje a analizar. Para ello, un camino es localizar los genes del core exclusivos de cada grupo de genomas y libres de parálogos. Aunque al momento existe publicada *metaphor*, una herramienta de selección de ortólogos [77], al comenzar este trabajo no existía un método disponible para ello. Finalmente una vez obtenidos los genes del core es deseable contar con un algoritmo que además los concatene y entregue un árbol filogenético.

## 0.5 La genómica comparativa como herramienta en la distinción de familias y enzimas promiscuas que participan en el metabolismo especializado.

Una parte del metabolismo especializado está compuesta por familias enzimáticas que evolucionaron de rutas de metabolismo central [78]. En las familias expandidas, ya sea por duplicación o por transferencia horizontal, las expansiones pueden retener la función química de las rutas centrales [79], así como también la función alternativa suele estar presente aún a bajos niveles antes de la divergencia o duplicación [33]. Por tanto las familias con expansiones en un linaje son candidatas a ser familias promiscuas en él. Se ha notado que en el linaje en que una familia enzimática es promiscua hay una zona de cambio en promiscuidad [39], Las expansiones de rutas centrales que participan en la síntesis de productos naturales son candidatos a presentar cambios en promiscuidad tanto a nivel familia como a nivel enzima.

por ello la observación de la retención de función ancestral al aparecer una función alternativa proporciona una zona favorable para la búsqueda de promiscuidad a nivel de enzima. En la familia existirá un gradiente de promiscuidad, las más cercanas a la zona de duplicación o divergencia tienen más posibilidades de tener un cambio en promiscuidad que las más conservadas y cercanas al metabolismo central.

Después de que en la sección anterior se estableció la posibilidad de mejorar las relaciones filogenéticas de un linaje, se abre la posibilidad de buscar en él expansiones de familias génicas de metabolismo central. Estas copias extra son candidatas a pertenecer a rutas de metabolismo especializado. Como prueba de concepto esta idea de minar genomas incorporando información evolutiva permitió la identificación de la biosíntesis de arsenolipidos [52]. La búsqueda de productos naturales cuenta entre sus premisas que estos se producen en vecindades genómicas llamadas clusters y que además clusters cercanos (ya sea en contenido genético o en la secuencia de sus componentes), exploran variaciones metabólicas, es decir sus enzimas catalizan reacciones sobre sustratos parecidos aunque no idénticos [52].

La primera versión 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.

1 ARTS [80] 2 EvoMining [52] Se ha avanzado en archaea [81]

Desarrollarla en combinación con algoritmos de búsqueda de cambios en la vecindad genómica la harán una plataforma ideal para abordar el problema de las familias, proporcionando una solución 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 genómica, flujo genético y dinámica molecular, como candidatos a reflejar la variación

en promiscuidad. Finalmente,

tomando como modelo biológico el phylum Actinobacteria, un grupo de bacterias reconocido por su diversidad metabólica donde se ha probado la existencia de promiscuidad enzimática.

Evomining es una plataforma bioinformática pensada para la identificación de productos naturales

Si se combinara evomining con la premisa de que vecindades distintas son marcadoras de funciones químicas distintas, al encontrar una familia expandida con vecindades genómicas diferentes se podría solventar la deficiencia de otros métodos bioinformáticos consistente en que para identificar familias promiscuas se debe conocer previamente un miembro promiscuo de la misma. (Fig 4) Así pues al combinar evomining con herramientas de vecindad genómica tanto de comparación como de visualización estaremos mejorando su funcionalidad en la identificación de familias promiscuas. En la siguiente sección BLA BKA

la variación es la materia prima de la evolución.

## 0.6 La genómica comparativa como herramienta en la priorización de clusters promiscuos

La promiscuidad nos interesa por su producción de variantes. Si bien en rutas centrales rescata la función en metabolismo secundario crea nuevas variantes moleculares que permiten adaptación, de hecho pangénomas grandes correlacionan con aparición de nuevas funciones enzimáticas. Considero que el concepto de promiscuidad puede ser extendido a un nuevo nivel. Promiscuidad enzimática, promiscuidad de familia, promiscuidad de cluster. En rutas centrales robustez, pero en metabolismo secundario platicidad. Pangéoma abierto cerrado, aquí proponemos organizar los clusters y finalmente la medida de su apertura. Variantes de clusters producen nuevos compuestos, ya sea por promiscuidad enzimática en un core conservado o por variación en la presencia / ausencia de genes. Un cluster recibe los mismos sustratos y los transforma en diferentes productos.

### 0.6.1 Expansion y contextos genómicos como herramienta de anotación funcional

Al evaluar la herramienta de análisis de promiscuidad PROMISE [36] en un set de datos de la familia HisA/PriA [39] obtuve que en su mejor desempeño es (huella molecular de tamaño 6) clasifica correctamente casi todas las no promiscuas, (HisA) pero no sucede lo mismo con la familia PriA donde tiene éxito 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 reflejando tal vez que en su conjunto

de entrenamiento no había miembros promiscuos ilvC. Por lo menos para estas familias el conjunto de entrenamiento o los descriptores no son suficientes para la anotacion de promiscuidad.

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 [1]. Existe evidencia de que cierto grado de promiscuidad o divergencia funcional precede a la duplicacion genica [62]. Por este motivo detectar expansions ya sea duplicaciones o transferencias horizontales [83], puede ser un buen punto de partida para determinar divergencia funcional y promiscuidad. No todas las expansions 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 [84], 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.

Para sistematizar el estudio de contextos y vecindades genomicas se desarrollo Search Tool for the Retrieval of Interacting Genes/Proteins STRING [85], 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 [86] 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 [44]. 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 distin-tivo 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 [41], (Juarez Vazquez et al 2015).

### 0.6.2 Contexto y vecindades genomicas

En 2012 fueron analizados 102 genomas de 29 familias de Actinobacteria [71]. sugiriendo que al menos en *Corynebacteria* el contexto y la vecindad genomica incidan en la sub-funcionalizacion de PriA en subHisA [41]. Respecto a IlvC, otra familia involucrada en la sintesis de aminoacidos fue estudiada y caracterizada bioquimicamente en 1 *Corynebacterium* y 8 *Streptomyces* [42]. 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 [42]

Natural products genomic era[88]

## 0.7 Estudio de una familia PriA

### 0.7.1 Caracterizacion in vivo

No todas las familias promiscuas provienen de expansiones, tal es el caso de PriA en Actinobacteria, donde no tiene expansiones y hasta el momento no se le conoce participacion en rutas de metabolismo especializado.

cuya promiscuidad es debida a la perdida de TrpF 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.

### 0.7.2 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).

### 0.7.3 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 [89]. 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 [91], 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[45]. 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 [39], 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 [71] se utilizo el metodo de dinamica molecular y se comparó el numero de confórmeros entre miembros de subHisA y PriA, resultando mayor el de PriA como corresponde a una enzima promiscua. El estudio sobre la relación dinámica-flexibilidad de  $\beta$ -lactamasas utiliza replica exchange, una variacion de dinamica molecular que corre replicas en paralelo a distintas temperaturas [92]. Una desventaja de este metodo es que por el costo computacional de las replicas agregar explicitamente otras moléculas 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 [13]. 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 sin embargo hasta ahora no se han reportado trabajos multidisciplinarios que involucren a todas las partes involucradas

# Pregunta biológica



# **Objetivos**

## **0.8 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)

## **0.9 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.



# Estrategias

**Obtener informacion genomica del phylum Actinobacteria.**

Colectar genomas de Actinobacteria de NCBI y de colecciones privadas.

## **0.9.1 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.

## **0.9.2 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.

**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.

**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 catalíticas, contexto genómico, vecindad genómica, número de conformeros e índice de flexibilidad.

**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 visualización e integrar la clasificación de vecindades genómicas como una herramienta adicional en la búsqueda de promiscuidad.

**Seleccionar miembros homólogos de la familia de enzimas.**

Se escogieron 41 Streptomyces repartidos en un árbol de rpoB de 400 Streptomyces con genoma disponible. Esta selección incluye los seis Streptomyces de los que se cuenta con cinética enzimática de PriA, tres de ellos con estructura cristalográfica.

**Medir cinéticas enzimáticas, contexto genómico, vecindad genómica, flexibilidad y número de conformeros.**

Determinar la pertenencia a uno de cuatro posibles contextos genómicos respecto al gen trpF. Estudiar la existencia de distintas vecindades genómicas. Determinar la cinética enzimática de 9 enzimas más buscando variabilidad en contexto genómico (sugeridas en la tabla 4). Obtener mediante una colaboración 37 modelos estructurales por homología y modelar dinámica molecular.

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

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

Número de conformeros e índice de promiscuidad.  
índice de flexibilidad y número de conformeros.

Número de conformeros y contexto genómico.  
índice de flexibilidad y contexto genómico.

Contexto genómico e índice de promiscuidad I.

Analizar las vecindades genómicas e índice de promiscuidad I.

**0.9.3 Desarrollar una metodologia para la detección *in vivo* de promiscuidad enzimática.**

Debido a cambios en flexibilidad o cambios de contexto genómico, se puede sospechar de diferencias en la función química de dos miembros de una familia de enzimas, sin conocer las diferencias a nivel de sustratos. Para investigar estos cambios *in vivo* se propone estudiar diferencias en perfiles metabolómicos de una colección de cepas en condiciones diversas.



# Metodologia

A continuacion describiré 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>.

## 0.9.4 La promiscuidad en familias enzimaticas.

### Actinobacteria genomica

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).

### **0.9.5 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 enzimáticas.

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) 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.p1](https://github.com/kbase/kbseed/blob/master/service-scripts/svr_CS.p1)). 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 CCi es un conjunto de contextos y ni es la cardinalidad de CCi Cl={CC1,CC2,...,CCk}

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

$$\#\$ \$ \quad M \neq \max_{i=1}^n \{n_i\} \quad i \in \{1, 2, \dots, k\} \quad n_i \neq M$$

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

M representa el contexto mas difundido de la enzima, dentro del grupo taxonomico considerado; mas 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  
 $S(Cl) = f(k, m, M) = c_1 k + c_2 m + c_3 M$

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.

### 0.9.6 Promiscuidad in vitro

#### Datos cineticos:

En todos los ensayos enzimaticos se busca medir una señal que permita una distincion clara entre sustrato y producto [93]. 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 [94]. 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 [67].

Finalmente se recopilaran datos cineticos de PriA tanto privados como los publicos reportados a la fecha en la Braunschweig ENzyme Database BRENDA [95]. 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 [12] que se define como:  $I = \frac{1}{N} \sum_{i=1}^N \frac{1}{K_i} \text{cat}_{\text{Kim}} / \text{cat}_{\text{Kicat}}$

#### 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, [96] para el modelado de la dinamica molecular. Esta parte del trabajo se realizara en colaboracion con el laboratorio de bioinformatica y biofisica computacional.

### 0.9.7 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 sustituira 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. [45] Posteriormente se combinan con espectrometria de masas que es una tecnica que mide el radio masa-carga de las partículas fragmentadas en iones. [45]. Los datos obtenidos de espectrometria de masas se procesaran utilizando redes moleculares, que consiste en agrupar los productos según la similitud de sus partes. Plan: 3 replicas tecnicas, 2 replicas biologicas de 5 cepas.

### 0.9.8 Consideraciones

Falsos negativos respecto a promiscuidad están muy extendidos en la literatura y en las bases de datos, en parte porque la mayoría 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 difícil 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 técnicas como el docking que está intimamente relacionado con la dinámica molecular [45]. Limitar el espacio de sustratos puede retroalimentarse con el estudio de la promiscuidad in vivo y viceversa.

Con los métodos propuestos en este trabajo solo se podrá detectar pérdida o ganancia de promiscuidad entre enzimas de organismos respecto a otros miembros dentro un grupo taxonómico, no así el estado de promiscuidad intrínseco a la enzima. Si dada una enzima no se detectan variaciones en contexto, vecindad genómica o flexibilidad dentro de un grupo taxonómico 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 parámetros descritos, con estos métodos no se puede sugerir promiscuidad. Es posible que al mirar en un grupo taxonómico más amplio se detecte una neofuncionalización de la familia aunque también es posible que exista una variable  $z$  como la flexibilidad de sustrato [20] que no se esté considerando y que explique o sea el mejor indicador para esta familia de promiscuidad enzimática.

Se debe considerar que si existe una correlación vecindad genómica-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[98]  
Archaeal core [99]  
Methanosarcina reconstruction [100]  
Archaea phylum[101]  
Prediction for possible products of promiscuous enzymes[102]  
Saxitoxin [103]  
Plants clusters [104]  
MiBIG [105]  
Metagenomics on Streptomyces [106]  
Sulfolobus reconstrucion [107]  
Archaeal Natural products[108]  
Computational Pangenomics [109]  
Cuántos genes “obtenidos por EvoMining” son core/ cloud/stand alone  
Qué porcentaje de genes únicos recupera EvoMining  
Eucarya paralogs reshape gene clusters [110]  
Microbial dark mater [111]  
Archaea anaerobica carbon [112] Archaea Eucarya gap loki[113]  
Archaea and eucarya[114]  
BPGA [115] genes esenciales bacteria minima[116]  
Radical [117]  
RaxML large phylogenies [118]  
R phylogenies [119]  
Streptomyces exploradores [120]  
LUCA [121] Luchando por el reconocimiento de Archaea[122],[123] The primary kingdoms [124]  
Prediccion aRchaeas [125]  
RASt archaea [126] Book Archaea [127]  
Computational methods for bacterial and archaeal genomes [128]  
Archaeas boook [129]  
GC content plasmido genoma [130] Genoma minimo[131]  
Phylogeny R [132]  
Cyanobacteria fluctuacion genomica y adaptacion [133]  
Ecology of cyanobacterua [134]  
Histidine biosynthesis[135]

- PriA reconstruction [136]
- Escala temporal bacterias [137]
- Pangenome size [138]
- variabilidad del 16s [139]



# Chapter 1

## Desarrollo de Orthocore e implementación de otras herramientas computacionales para entender el pangenoma de un linaje genómico.

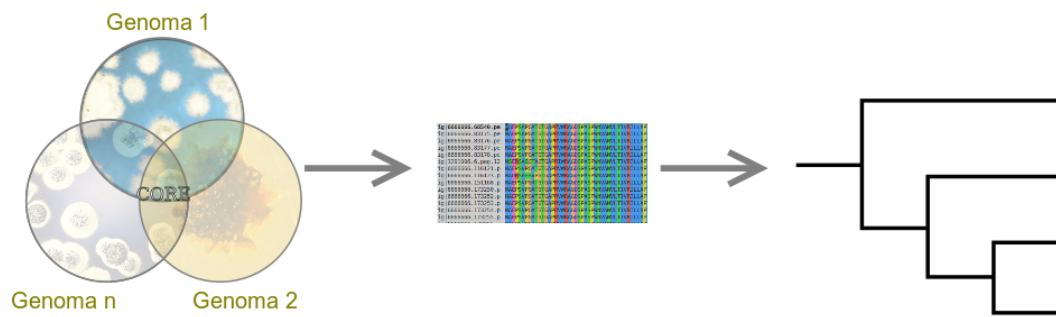


Figure 1.1: Orthocore calcula las familias génicas comunes de un linaje genómico. Después de un proceso de filtrado, alineamiento y curación concatena estas familias y entrega una reconstrucción filogenética.

El pangenoma es el contenido génico total de un linaje taxonómico. Las familias génicas de un pangenoma pueden clasificarse según sus patrones de presencia-ausencia en cada genoma del linaje. De acuerdo a esta clasificación los principales grupos de familias génicas en un pangenoma son el *core*, el *shell* y el *cloud* (dispensable) genome. El *core genome* es el conjunto de familias con presencia en todos los genomas del linaje. Por ejemplo, la secuencia de la subunidad 16s del gene rRNA, así como diversos genes ribosomales que suelen estar en el core de la gran mayoría de los linajes bacterianos.

El *shell genome* es el grupo de familias presentes en la mayoría de los genomas pero no en todos. En el *shell* se ubican por ejemplo familias que estaban en el *core genome* pero que algunas bacterias del linaje sufrieron una dinámica de pérdida (o ganancia) génica. Mientras que el *cloud genome* o dispensable genome es aquel grupo de familias que sólo ocurre en unos cuantos genomas del linaje.

La organización filogenética de un linaje genómico permite la observación de pérdida y ganancia de familias génicas en organismos cercanos. Si los organismos están desordenados es difícil apreciar la dinámica genómica de aparición-desaparición de miembros de una familia génica. Ordenar los genomas de un linaje facilita apreciar cambios en el número de copias de una familia. Esto es relevante en el marco de esta tesis ya que cambios en los perfiles de promiscuidad podrían estar relacionados a copias extras de organismos cercanos. Orthocore es un algoritmo que utiliza el core conservado de un linaje genómico para facilitar la organización filogenética de sus organismos Figure 1.1

## 1.1 La distribución de la función metabólica de las familias del pangenoma depende de la variabilidad del linaje seleccionado.

El número de familias génicas presentes en el pangenoma, así como su distribución en el *core*, *shell* y *dispensable genome* depende de la elección de los genomas y del linaje genómico. Para entender esto se puede pensar en un ejemplo extremo, consideremos una bacteria de 1000 familias de genes de la cual se obtienen secuencias de diez genomas de la misma cepa. Estas secuenciaciones deberían ser prácticamente idénticas y en ese caso el *core genome* sería 1000 familias, el *shell* y el *dispensable genome* serían cero. En este caso, todo el metabolismo, tanto el central como el especializado estarían conservados dentro del *core genome*, ya que el *shell* y el *dispensable genome* se encuentran vacíos. Sin embargo, si variamos el linaje taxonómico, y ahora estudiamos el pangenoma de 10 especies distintas del género *Streptomyces* ahora el core genome estará compuesto por aproximadamente un tercio de su tamaño promedio, y dentro del *core genome* es donde se encontrarán muchos de las familias dedicadas al metabolismo central o conservado (por ejemplo, familias de la glicólisis o síntesis de aminoácidos). En cambio muchas de las familias dedicadas al metabolismo especializado y pertenecientes a clústers biosintéticos de productos naturales (BGCs) estarán en el *dispensable genome* pues *Streptomyces* es productor de una gran variedad de metabolitos especializados y cada especie suele tener su producto característico. Figure 1.2

El *core genome* de un linaje, además de tener familias conservadas y prácticamente presentes en todo el dominio Bacteria también puede contener familias marcadoras Figure 1.3 . Estos genes marcadores permiten realizar pruebas de diagnóstico para colonizaciones bacterianas. A las familias que están presentes en el *core genome* de un linaje A, pero que están completamente ausentes de un linaje B se les llama

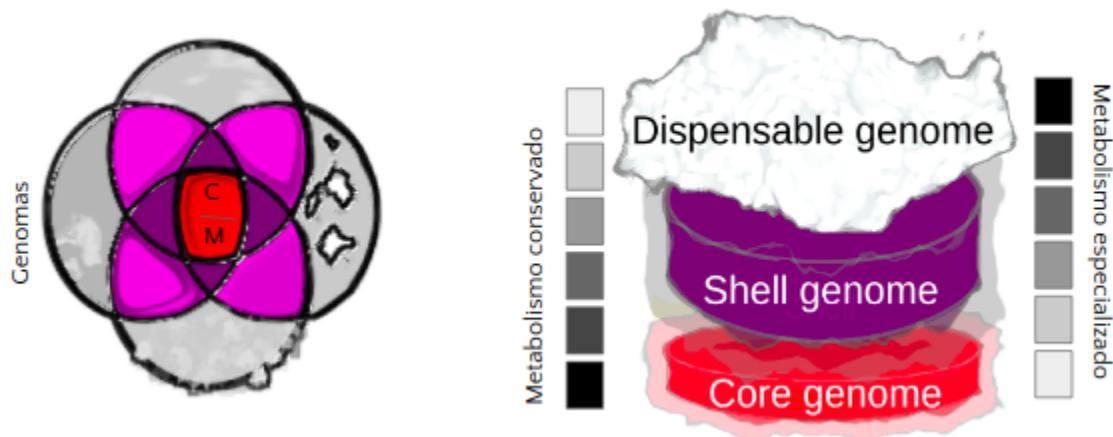


Figure 1.2: El pangenoma de un conjunto de genomas de un linaje puede ser clasificado en varios grupos. En este ejemplo, en el lado izquierdo de la figura se observa en gris el *genoma dispensable* compuesto por familias génicas presentes sólo en un genoma. En dos tonos de morado observamos el *shell genome*, familias que están presentes en la mayoría de los genomas del linaje, en este caso dos o tres genomas. Finalmente en rojo se muestra el *core*, aquellas familias presentes en todos los genomas del linaje. El *core* contiene tanto familias muy conservadas con una sola copia por genoma ( C ), como familias expandidas. Las familias marcadoras ( M ) pueden ser parte del *core* conservado o de las familias expandidas. Del lado derecho se muestra una representación del pangenoma para cualquier número de genomas. Familias de metabolismo conservado tenderán a estar concentradas entre el *core* y el *shell genome*, mientras que el metabolismo especializado tendrá más representantes en el dispensable que en el *core genome*. Sin embargo tanto el tipo de metabolismo como el tamaño del *core*, *shell* y *genoma dispensable* pueden variar según la diversidad de los organismos seleccionados.

marcadoras. Por ejemplo genes conservados en la especie *Streptomyces coelicolor* pero no conservados en *Streptomyces rimosus* son genes marcadores de *Streptomyces coelicolor* respecto de *Streptomyces rimosus*. Estos mismos marcadores tal vez no sean marcadores respecto de *Streptomyces lividans*, a pesar de la cercanía taxonómica entre estos organismos. La presencia de genes marcadores en el core depende de ambos linajes, por lo que es importante contar con algoritmos que permitan automatizar su cálculo.

El número de familias en el pangenoma, ya sea en el *core*, *shell* o *dispensable* genome no sólo depende de la divergencia o proximidad taxonómica de los organismos del linaje seleccionado, también depende de lo variable que sea el contenido génico en los genomas del linaje. A esta característica se le conoce como apertura. Hay especies, por ejemplo algunos patógenos, cuyo pangenoma se encuentra sumamente cerrado en el sentido de que no importa cuántos genomas se agreguen, el número de familias parece converger y ser asintótico rápidamente a una cota superior. En cambio especies o géneros que viven en una gran diversidad de hábitats suelen tener un pangenoma abierto. Esto significa que cada vez que se agrega un nuevo genoma aparecen otras familias que no estaban en los genomas anteriores. En los linajes con pangenoma abierto el número de familias nuevas al agregar un genomas seguirá una tendencia creciente y no asintótica.

Además de la apertura, existen otros intentos de cuantificar la diversidad génica de un linaje. Está por ejemplo la fluidez, definida como el promedio de familias únicas entre familias totales por pares de genomas. El pangenoma bacteriano total, es decir el total de familias génicas en el dominio Bacteria es considerado abierto.

Finalmente la distribución de las funciones metabólicas encontrada en los subconjuntos del pangenoma (*core*, *shell* y *dispensable* genome) está relacionada a la proximidad filogenética de los organismos seleccionados en el estudio. Entre más diversos sean los organismos menos familias dedicadas exclusivamente a metabolismo especializado abundarán en el *core/shell genome*. La diversidad provocará que lo único que tengan los genomas de estos organismos en común sean funciones conservadas por una amplia variedad de especies bacterianas. Ahora bien, muchas familias de metabolismo especializado provienen de reclutamientos de copias extra de familias de metabolismo conservado. Así pues aunque decrezca el número de familias con exclusividad en metabolismo especializado en el *core y shell genome*, estos subconjuntos del pangenoma aún pueden contener familias conservadas que tengan copias extra en proceso de reclutamiento para algún Cluster biosintético de genes (BGCs) de metabolismo especializado. Considerando las reflexiones anteriores, entre más diverso sea un linaje, más tenderá su *core genome* a contener exclusivamente familias de metabolismo conservado mientras que su *dispensable genome* estará formado mayormente por familias de enzimas del metabolismo especializado.

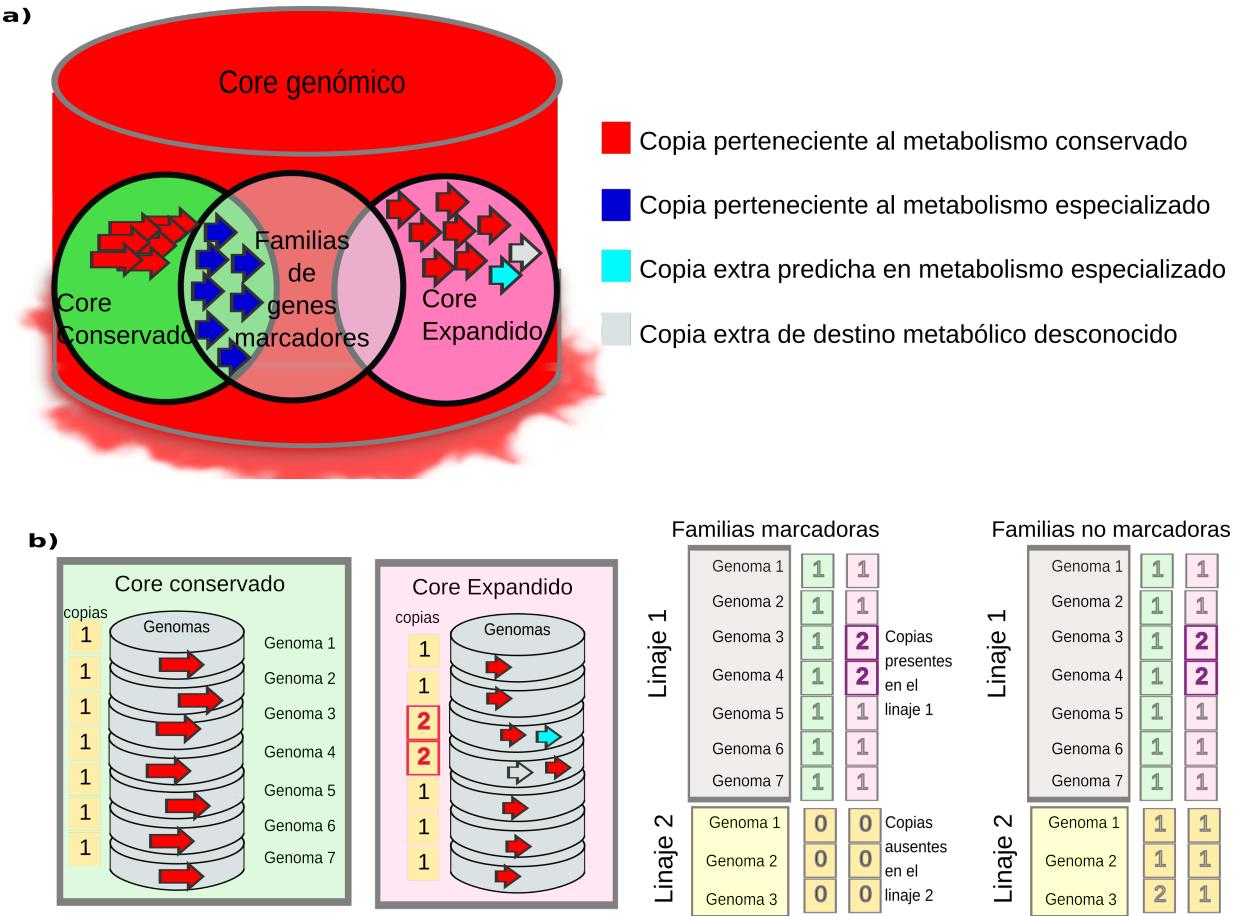


Figure 1.3: El core genome puede contener familias con funciones en distintos grupos metabólicos así como diversidad en el número de copias. Arriba se muestra que en el core pueden coexistir familias tanto de copia única como con expansiones. Las familias con funciones en el metabolismo conservado suelen concentrarse en el *core genome* (rojas), pero también dependiendo de los organismos seleccionados pueden encontrarse ya sea familias enteras o algunas copias dedicadas al metabolismo especializado (azul). No de todas las copias se conocerá su función, algunas pueden tener un destino metabólico desconocido (gris) o bien ser predichas por algún algoritmo como parte del metabolismo especializado (cyan). Abajo a la izquierda se comparan familias del *core conservado* con exactamente una copia por genoma contra familias del *core expandido*. Ambas pertenecen al *core*, pero en el *core expandido* hay dos genomas que tienen una copia extra en esta familia, una cyan y una gris, que podría dificultar la elección de los verdaderos ortólogos. A la derecha se ejemplifican familias de genes marcadores, útiles para identificar un linaje genómico. Tanto familias del *core conservado* como del *core expandido* pueden ser familias marcadores, siempre que exista al menos una copia de cada familia en el linaje 1 y ninguna copia en el linaje 2. Las familias dejan de ser marcadores cuando el linaje 2 contiene al menos una copia en algún genoma.

## 1.2 El core conservado permite la reconstrucción de filogenias complicadas

Orthocore es el desarrollo bioinformático que realicé para calcular las familias génicas más conservadas del *core genome*. Dos genes son homólogos si poseen un ancestro común, entre los principales grupos de homólogos están ortólogos y parálogos. Los ortólogos provienen de eventos de especiación de un ancestro común mientras que los parálogos evolucionan por eventos de duplicación. Orthocore obtiene un subconjunto del *core genome*: el *core conservado*, es decir, familias de ortólogos presentes en todos los genomas del grupo y que además son libres de parálogos de difícil identificación. El *core conservado* facilita la organización en árboles filogenéticos de organismos de un linaje genómico.

La comparación de la variación molecular entre ortólogos ha sido utilizada para establecer relaciones filogenéticas entre organismos. Esta técnica ha dado lugar a grandes descubrimientos. Por ejemplo comparar la secuencia de la subunidad 16s del gene RNA ribosomal condujo a Woese al descubrimiento del dominio Archaea en 1977 [124]. Un árbol de especies suele hacerse con secuencias de familias que pertenecen al core genome de un Dominio, por ejemplo las familias 16s o RpoB en los Dominios Bacteria y Archaea. Algunos autores realizan árboles multilocus para mejorar la resolución de árboles de especies realizados mediante la comparación de secuencias de 16s. Los genes seleccionados para los árboles multilocus deben estar en todos los organismos y no tener copias extra tan parecidas que puedan confundirse y entorpecer la reconstrucción filogenética, es decir, las familias seleccionadas deben ser parte del *core conservado*. Orthocore automatiza la identificación de estas familias.

Entre los factores importantes para establecer las relaciones filogenéticas diferenciando entre Archaea y Bacteria están los siguientes: 1) la presencia conservada de la subunidad de 16s en los tres dominios mencionados, y 2) la suficiente divergencia entre estas secuencias en los organismos de dichos dominios. Ahora bien, establecer relaciones filogenéticas entre Archaea y Bacteria es en cierto sentido más sencillo que establecerlas entre organismos pertenecientes al mismo género o inclusive a la misma especie. En ocasiones, como en el caso del género *Streptomyces*, la secuencia de 16s por sí sola no posee la suficiente variación para resolver la filogenia [140]. En *Streptomyces* la variación entre estas secuencias suele ser menor al 1%. Para resolver el problema de escasa variación en secuencias de 16s se pueden concatenar las secuencias de otros ortólogos, siempre que estos aparezcan en todos los organismos que se estén estudiando, es decir, siempre que sean parte del core genómico.

## 1.3 El algoritmo de Orthocore

Los ortólogos suelen identificarse por similitud de secuencia, pero si se realiza la identificación manualmente también se suelen capturar parálogos que pueden confundir

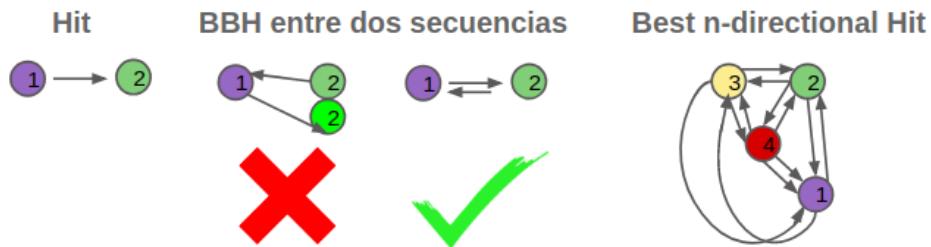


Figure 1.4: Orthocore utiliza los mejores hits n-direccionales para obtener grupos de ortólogos. Un hit es el mejor resultado de una secuencia en otro genoma. Un Bidireccional Best Hit es el mejor hit bidireccional. La secuencia 2 es el mejor hit de la secuencia 1 en el genoma 2 y recíprocamente, la secuencia 1 es el mejor hit de la secuencia 2 en el genoma 1. La existencia de una copia extra muy parecida a la secuencia 2 puede romper el BBH. Un mejor hit n-direccional debe ser BBH todos contra todos garantizando que estas secuencias están muy conservadas entre sí.

la elucidación de eventos de especiación. Orthocore automatizó la búsqueda de ortólogos y el filtrado de parálogos en genomas procariontes mediante la generalización de la definición del mejor hit bidireccional (BBH por sus siglas en inglés) Figure 1.4 . Dos secuencias son BBH si cada una es el mejor hit de un algoritmo de distancia (BLAST usualmente) en el genoma de origen de la otra. Una primera generalización para obtener el set de ortólogos de una familia del core es definir un genoma de referencia y tomar los BBH respecto a ese genoma. En la práctica, esta definición da como resultado distintos resultados según el genoma de referencia, haciendo que algunos parálogos no sea filtrados.

Para solventar esta dificultad se definió en Orthocore el concepto de mejores hits multidireccionales. Un conjunto de genes son mejores hits multidireccionales si todos entre sí son BBH por pares. Es decir si cada gen fuera un punto y ser mejor hit se expresara como una conexión con dirección todos los puntos estarían conectados por una flecha de ida y otra de regreso. Con este método se eliminó la dependencia de un genoma de referencia. Esta restricción también ocasiona que en grupos muy grandes por ejemplo más de 100 genomas de distintas especies, o muy diversos de distintos dominios, o muy fragmentados como con contigs de en promedio 3 Mbp, el core conservado puede quedar vacío.

### 1.3.1 Componentes técnicos de Orthocore

Orthocore es un tubería escrita en perl que incorpora los hits multidireccionales permitiendo obtener y usar el core conservado para realizar una reconstrucción filogenética mediante los siguientes pasos:

- Obtiene el core conservado: los mejores n-direccional hits (blastp).
- Alinea cada familia del core conservado (MUSCLE).
- Cura automáticamente cada familia del core conservado (Gblocks).

-Concatena las familias del core conservado formando una matriz de aminoácidos.

-Realiza una reconstrucción filogenética de la matriz de aminoácidos (FastTree)

-Provee las funciones del core conservado según la anotación funcional de RAST.

Existen otros algoritmos como OrthoMCL, y fastOrtho que dividen pangenomas en clusters de familias de genes, get\_homologues y metaphor que obtienen el core y filtran buscando verdaderas relaciones de homología, y finalmente BPGA que hace reconstrucciones filogenéticas tanto según el core como según el pangenoma. Sin embargo Orthocore resolvió en su momento la necesidad específica de proporcionar una matriz concatenada de genes del core conservado lista para utilizarse en un árbol multilocus. Adicionalmente, como Orthocore fue diseñado para trabajar con la anotación de la plataforma RAST, también se obtiene la anotación funcional tanto de familias del core como del complemento.

Orthocore incorpora todas las dependencias en un contenedor de docker disponible en <https://github.com/nselem/orthocore>. Además en este contenedor está un script que permite bajar genomas de NCBI masivamente y posteriormente anotarlos en RAST desde la terminal. Los protocolos de uso se encuentran al final de este capítulo.

### 1.3.2 Aplicaciones de Orthocore, identificación del core conservado y de familias de genes marcadores.

Cuatro aplicaciones de Orthocore serán presentadas en las siguientes secciones de este capítulo. En la primera aplicación el core conservado en *Actinomycetales* permitió organizar filogenéticamente a este orden. Esta organización facilitó el entendimiento en cambios de promiscuidad de la familia enzimática PriA mediante la distinción de patrones de pérdida y ganancia de genes en las rutas de síntesis de histidina y triptófano. En la segunda aplicación cepas de *Salmonella tifi* fueron ordenadas filogenéticamente. La tercera aplicación permitió realizar una reconstrucción filogenética del orden *Nostocales* del phylum *Cyanobacteria* y comparar así patrones de presencia y ausencia de clusters de genes biosintéticos. Finalmente, en organismos del microbioma del tomate orthocore de utilizó para identificar genes marcadores que permitieran distinguir cepas de *Clavibacter Michiganensis* de otras especies de *Micrococcaceae*.

Perfiles de promiscuidad de PriA en el orden *Actinomycetales* se relacionan con la especiación. Para mejorar el entendimiento de los cambios en promiscuidad enzimática de PriA, se necesitaba entender filogenéticamente al orden *Actinomycetales*, un camino era obtener su core conservado. Orthocore fue diseñado para resolver este problema. Con el resultado de Orthocore se realizó un árbol de especies donde se observaron patrones de pérdida y ganancia de genes en la vecindad genómica del gen que codifica para PriA. Se encontró que hay clados de *Actinomyces* donde los genes correspondientes a la síntesis de histidina no estaban en la vecindad genómica de PriA, y mediante la realización de cinéticas enzimáticas se comprobó que la actividad de catalizar la reacción correspondiente a HisA estaba perdida en estos organismos. A estas enzimas se les

llamó subTrpF ya que sólo poseían la capacidad de catalizar la reacción correspondiente a la familia TrpF. Del mismo modo existían clados que perdieron los genes de síntesis de triptófano en la vecindad de PriA y estas enzimas se subfuncionalizaron a la familia subHisA. De estos datos se observa que en estos organismos la especiación coincidió con el cambio de promiscuidad en la familia PriA, acorde a la pérdida y ganancia de genes vecinos. La promiscuidad puede co-ocurrir con variaciones en el contexto genómico, pudiendo estos cambios ser una marca para sugerir cambio funcional en una familia.

### 1.3.3 Genes de islas de patogenicidad de *Salmonella* en México están conservados en la mayoría de los genomas.

Para estudiar la diversidad de *Salmonella* presente en alimentos en México, primero se ensamblaron y anotaron genomas secuenciados para el trabajo “distribución de los genes de la toxina VirB/D4 en plásmidos de bovino asociados a *Salmonella* no tifoídil”. Los genomas fueron ensamblados en Patrick y se desarrolló myRAST, una tubería previa a Orthocore para la anotación automática de genomas ensamblados en RAST. El protocolo de myRAST puede encontrarse al final de este capítulo.

Orthocore fue usado aquí para reconstruir filogenias de *Salmonella* y además fue integrado como parte de CORASON, el cual se reporta en el Capítulo 3, es el algoritmo que sirve para visualizar variantes de clusters organizados filogenéticamente. Los clusters de genes pueden ser biosintéticos, islas de patogenicidad, operones o cualquier región parcialmente sintética de un genoma bacteriano centrada en un gen. En este caso se observó una alta conservación de toxinas tifoidales en islas de patogenicidad de *Salmonella*. Éstas fueron identificadas en 76% de las cepas analizadas y posteriormente visualizadas mediante CORASON.

### 1.3.4 Nostoc provenientes del metagenoma de cíadas se agrupan Cyanobacteria

Las Cyanobacterias son un phylum de bacterias que se han adaptado a diversos ambientes. Aunque muchas de ellas son marinas algunas Cyanobacterias viven como simbiontes de plantas. En particular las cíadas han desarrollado un tipo especial de raíz donde se sabe que vive como simbionte el género *Nostoc*. La presencia de *Nostoc* en la raíz coraloides de las cíadas es fácilmente distinguible por la formación de un anillo verde conocido como anillo cyanobacterial. En la Figure 1.5 se muestra la filogenia de 76 Cyanobacterias de 7 órdenes distintos construida con 198 proteínas del core conservado obtenidas por Orthocore. En esta reconstrucción se puede observar que los *Nostoc* asociados a plantas tienden a agruparse en la filogenia.

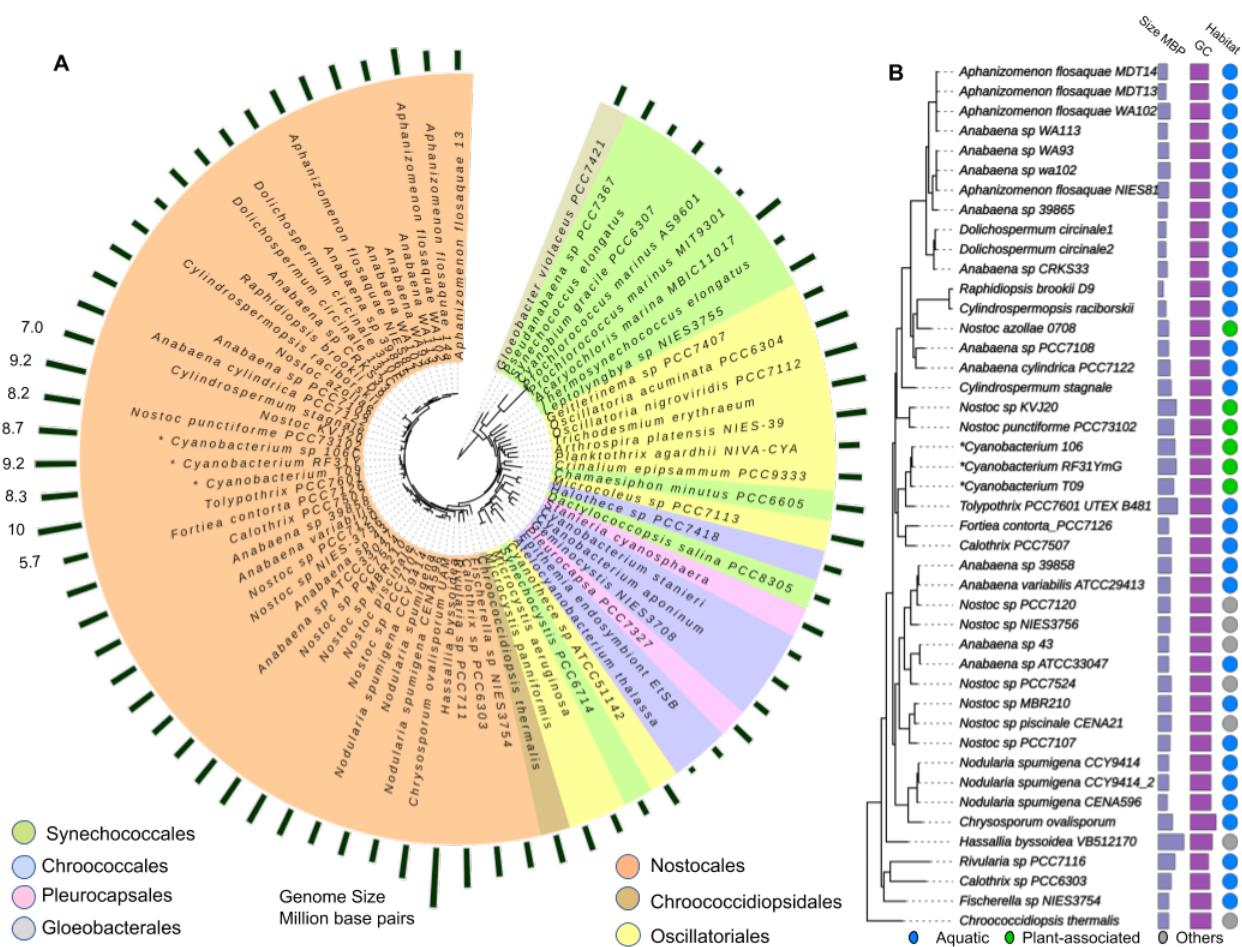


Figure 1.5: Reconstrucción de 76 taxa provenientes de 7 órdenes de Cyanobacteria. La matriz final incluyó 45,475 aminoácidos curados de 198 familias de proteínas pertenecientes al *core conservado*. A la derecha se muestra un zoom sobre el orden *Nostocales*. En este orden se incluyen algunas bacterias simbiontes de cíadas. Metadatos como el tamaño de genoma, contenido de GC y hábitat de origen muestran una posible tendencia de incremento de tamaño en los genomas provenientes del microbioma de plantas.

### 1.3.5 Identificación de genes marcadores de *Clavibacter michiganensis*

*Micrococcales* es un orden de Actinobacteria que contiene a *Clavibacter*, *Micrococcus* y *Microbacterium*, entre otros microorganismos. El género *Clavibacter* comprende especies que pueden causar enfermedades en diversas plantas. En particular la especie *Clavibacter michiganensis* es una bacteria causante de la enfermedad del cáncer del tomate. *Clavibacter michiganensis* ha sido frecuentemente aislada en compañía de otros *Micrococcales* morfológicamente parecidos. La distinción entre microorganismos debida a la comparación de la secuencia de 16S no era suficiente para distinguir entre *Micrococcales* del microbioma del tomate, por lo que una prueba de diagnóstico se hacía necesaria. Se habían utilizado como marcadores genes como *tomA*, *ppaC* y *celA* entre otros, sin embargo estas elecciones en ocasiones resultaban en falsos positivos según árboles de especies de 16S, por lo que nuevos marcadores eran necesarios.

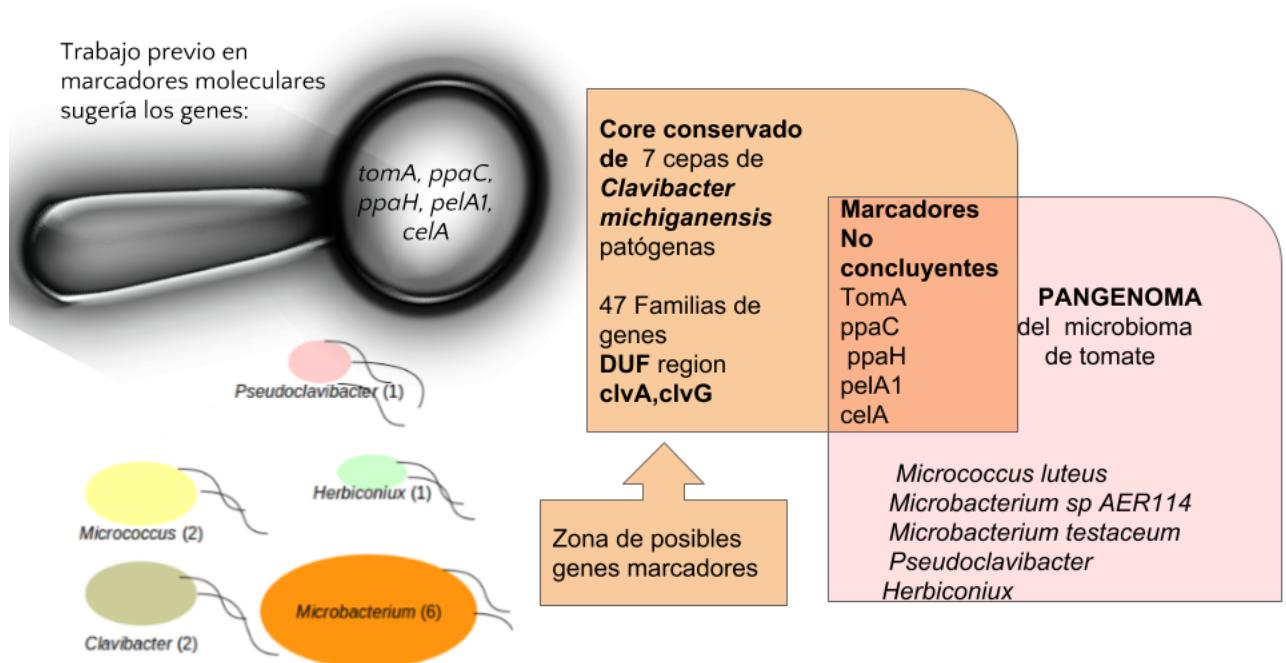


Figure 1.6: Los marcadores moleculares previos a este trabajo no permitían diferenciar correctamente a *Clavibacter michiganensis* respecto al microbioma del tomate en invernaderos mexicanos. Arriba a la izquierda se muestran los antiguos marcadores *tomA*, *ppaC*, *ppaH*, *pelA1*, *celA*. Abajo organismos pertenecientes al microbioma del tomate. Orthocore obtuvo el core conservado de siete cepas de *Cmm* patógenas, y este core se utilizó para definir nuevos marcadores. Aunque *tomA* pertenecía al core conservado de *Cmm*, estaba también incluido en el pangenoma del microbioma del tomate. Después de calcular la intersección core conservado de *Cmm* y pangenoma del microbioma se obtuvieron entre las familias marcadoras genes *clv* parte del cluster biosintético de clavidicina (michiganina).

Al analizar en RAST genomas de *Microbacterium* y *Micrococcus* aislados de tomate

se encontró que en efecto, *tomA* y los otros marcadores propuestos previamente no eran exclusivos de *Clavibacter michiganensis* (*Cmm*). Al utilizar Orthocore en siete genomas de *Cmm* encontramos que varios genes del cluster biosintético de michiganina (BGC0000528 en MIBiG) codificado por los genes *clvAFGLKM* pertenecían al core conservado, pero que al agregar los genomas no *Cmm* del resto del microbioma del tomate los genes *clv* se pierden. El descubrimiento de que *clv* pertenecía al core de *Cmm* se realizó con secuencias de genomas muy fragmentados, en la Figure 1.6 se muestran las cepas originales que fueron analizadas.

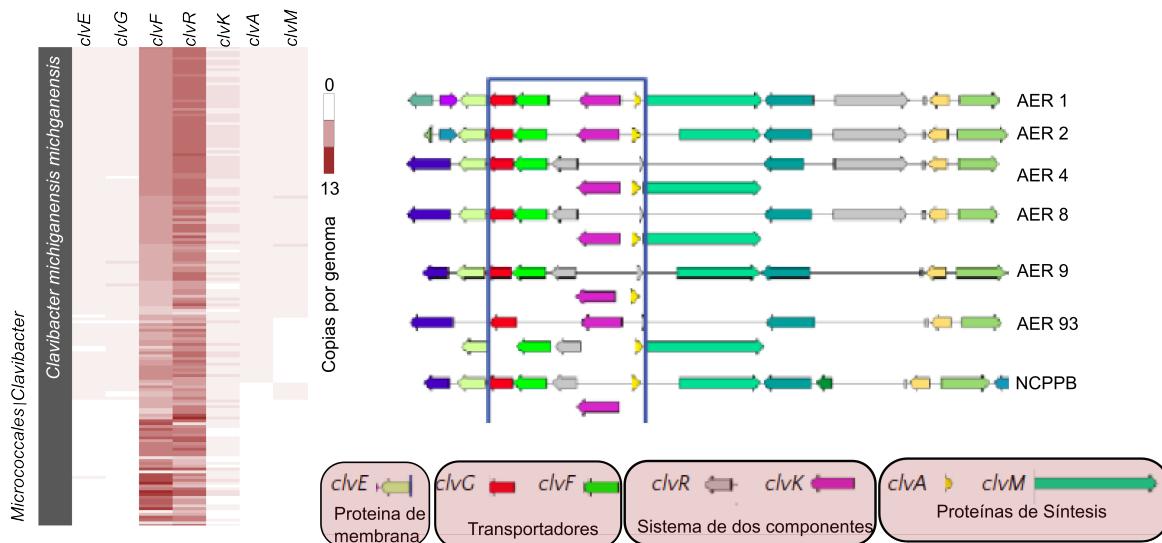


Figure 1.7: clavibacter

Esta observación se corroboró con más genomas, Figure 1.7 en este caso se muestran como ejemplo 10 genomas de bacterias del microbioma del tomate, entre ellas siete *Clavibacter*, seis de tomates de invernadero y uno *Clavibacter* proveniente de tomate silvestre, *Clavibacter RA1B* que fue reportado en la tesis de maestría de Yanez. Con Orthocore vemos que el tamaño del core decrece al ir agregando genomas de *Clavibacter* y decrece aún más rápido al agregar los genomas de *Micrococcus* y *Microbacterium*. La reconstrucción filogenética de este microbioma, ubica a *Clavibacter RA1B* cerca de los otros *Cmm* pero no en un clado junto con ellos. Una búsqueda por blast revela que los genes marcadores de *Cmm*: *clvF*, *clvR* son también marcadores de *Clavibacter RA1B* pero no así *clvA* y *clvG* que solamente están presentes en el core de *Cmm*. Sin embargo *clvF* y *clvR* no están en el contexto del cluster de michiganina en RA1B y su similitud de secuencia es menor que la que se observa entre los otros *Cmm*.

De hecho al considerar más genomas dentro del microbioma del tomate, la familia *clvF* no solo no está en el core conservado de *Microbacterium* y *Micrococcus*, sino que no está presente en ningún otro genoma distinto a *Clavibacter*. Con distintos niveles de conservación de secuencia los genes *clv* son un buen marcador para distinguir *Cmm* de otras especies, por esta razón estos genes aún se encuentran en uso como genes marcadores de *Cmm*. Esta misma conclusión se reporta en el trabajo de yasuuhara,

su observación de que las familias *clvA*, *clvF* y *clvG* son exclusivas de *Cmm* es independiente de la nuestra, fue hecha sin análisis genómicos y basada en evidencia experimental [141]. Este descubrimiento ha permitido bajar los costos de identificación de *Clavibacter*, ya que ahora en lugar de enviar a secuenciar el genoma es suficiente identificar por PCR *clvF* en conjunto con otro marcadores.

Con Orthocore además de obtener los genes marcadores podemos obtener también la matriz del core conservado para realizar la reconstrucción filogenética de especies cercanas de *Clavibacter*. Debido a que es importante para los agricultores conocer de dónde provienen las bacterias que infectan al tomate, una de las líneas de investigación del laboratorio de evolución de la diversidad metabólica, es la organización taxonómica de cepas de *Cmm* y de otras bacterias del microbioma de tomate. Por este motivo se tienen unos doscientos genomas privados y se esperan más a futuro. Orthocore colabora con esta investigación proporcionando matrices multilocus que pueden diferenciar entre *Clavibacter* de la misma o de diferente especie.

Debido al intercambio génico por transferencia horizontal en las bacterias, es posible que los genes marcadores actuales alguna vez aparezcan en otros organismos. También las bacterias pierden continuamente genes, por lo que es posible que algún gen marcador de *Cmm* se pierda en una cierta cepa Figure 1.8 . Esto hace que la definición de estar presentes en el core del grupo de interés y ausentes totalmente de cada uno de los genomas de otro linaje ya no funcione completamente. Sin embargo, en estas dos situaciones presentadas, ganancia de genes marcadores de organismos externos al linaje original o pérdida de genes marcadores en algunas cepas, se sigue cumpliendo que los ex genes marcadores, estarán presentes en la mayoría de los organismos del linaje de interés y ausentes de la mayoría de los organismos del linaje externo. Por ello, se pensó que está definición de genes marcadores se podía generalizar clasificando a grupos de genes ortólogos acorde a sus porcentajes de ocurrencia. Esta idea se desarrolló en la herramienta clavisual, explicada en la siguiente sección.

## 1.4 Clavisual: Identificación de genes marcadores a un cierto porcentaje de grupos seleccionados

La idea de que Orthocore puede ser usado para obtener los genes marcadores de un grupo taxonómico frente a otro fue generalizada en el backend del software Clavisual. Ya se ha explicado previamente que el core puede salir vacío por diversas razones, entre ellas baja calidad de los genomas, o que éstos provengan de organismos muy divergentes, verdaderas razones biológicas como dinámica génica o un core no convergente. Así pues, es posible que si sólo se utiliza el core no se obtengan marcadores. Pero el core puede relajarse de varias maneras una de ellas es el Pseudocore, donde en lugar de multidireccional hits se toman BBH a un genoma de referencia. Otra forma es establecer un porcentaje de presencia /ausencia de interés.

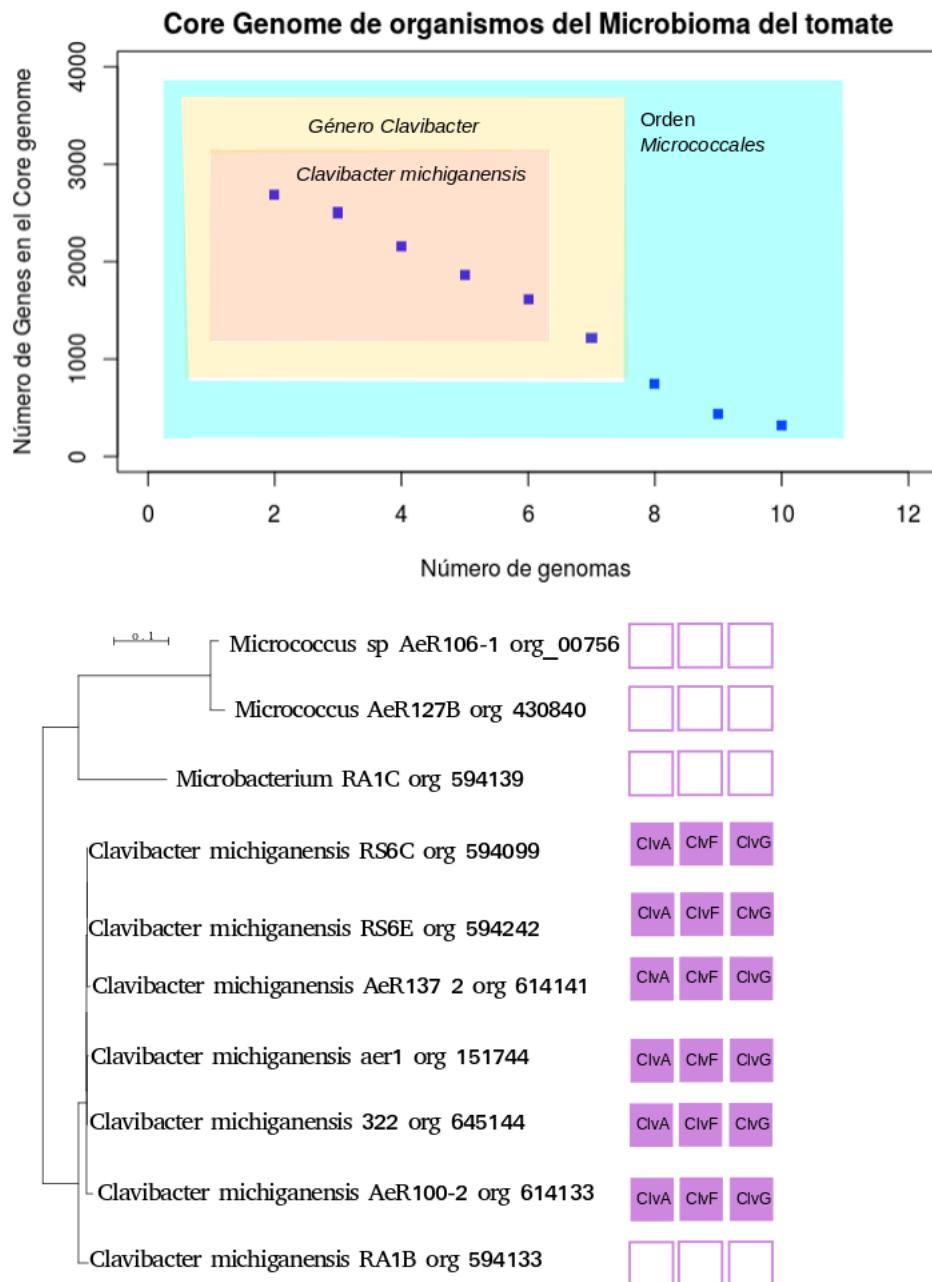


Figure 1.8: Aquí vemos los marcadores de EEvomining

El pseudocore consiste en \_\_\_\_\_ y la metodología está depositada en github en el repositorio\_\_\_\_\_. El blast fue optimizado cambiando hacer un blast todos contra todos por archivos genómicos individuales `genomai_vs_genomaj.blast` que luego son concatenandos según se necesiten.

Los porcentajes de genomas son diferentes porque al no bastar los best bidireccional hits conservados, todo el pangenoma es decir todos los genes contenidos en los genomas del grupo de interés necesitan ser clasificados por familias, para de ahí obtener las familias que tienen presencia en un porcentaje %p y ausencia en un porcentaje a% del grupo externo. Estos perfiles fueron desarrollados para Clavisual Figure 1.9 utilizando FastOrtho para clasificar las familias y de ahí obtener los grupos. Con ellos se consiguieron marcadores para *Kurtobacterium*.

Finalmente Clavisual despliega un árbol realizado con el Pseudocore respecto a un conjunto de genes de Cmm NCPP previamente seleccionados. En este árbol clavisual permite la visualización de metadatos, como año, género de la bacteria, estado de salud de la planta e invernadero donde fue aislada.

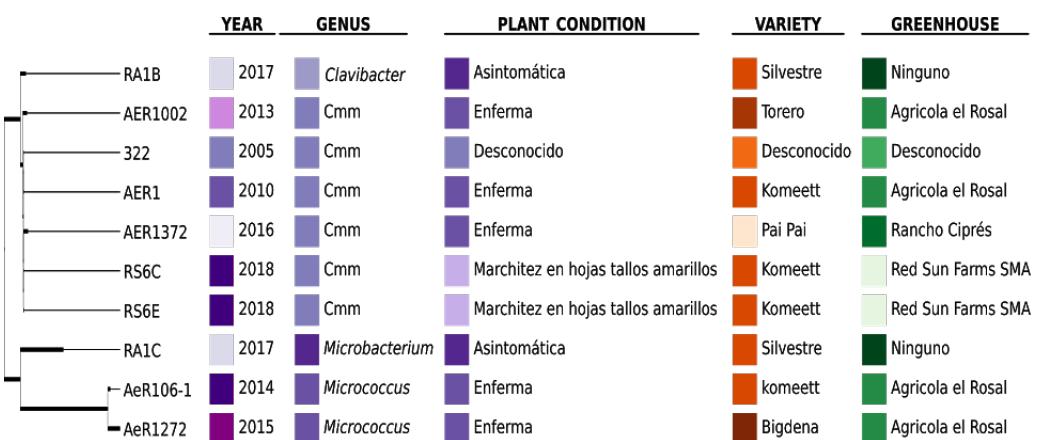


Figure 1.9: Aquí vemos un arbol

#### 1.4.1 El pangenoma de *Clavibacter Michiganensis* es abierto

Después de desarrollar métodos de identificación de genes marcadores y generalizarlo a obtener grupos con patrones de presencia/ausencia definidos por el usuario, quedaba por responder la pregunta cómo es el pangenoma de *Cmm*. Algunos autores consideran que el pangenoma de patógenos es reducido porque sus genomas suelen sufrir proceso de reducción de tamaño debido a la pérdida de genes. Como *Cmm* es un patógeno de planta quedaba por investigar cómo es su pangenoma. ¿Es posible saturar el contenido génico de *Cmm* con sólo secuenciar más genomas?. Aunque actualmente existen ya herramientas web para el análisis de pangenoma, en su momento se utilizó el software bpga que se corre desde la terminal. Para facilitar su instalación se desarrolló un docker que se describe más tarde en este capítulo en la sección de las descripciones

técnicas. Como ejemplo de su funcionamiento, se analizó el pangenoma de los mismos diez genomas del pangenoma del tomate utilizados en la visualización de clavvisual Figure 1.10 .

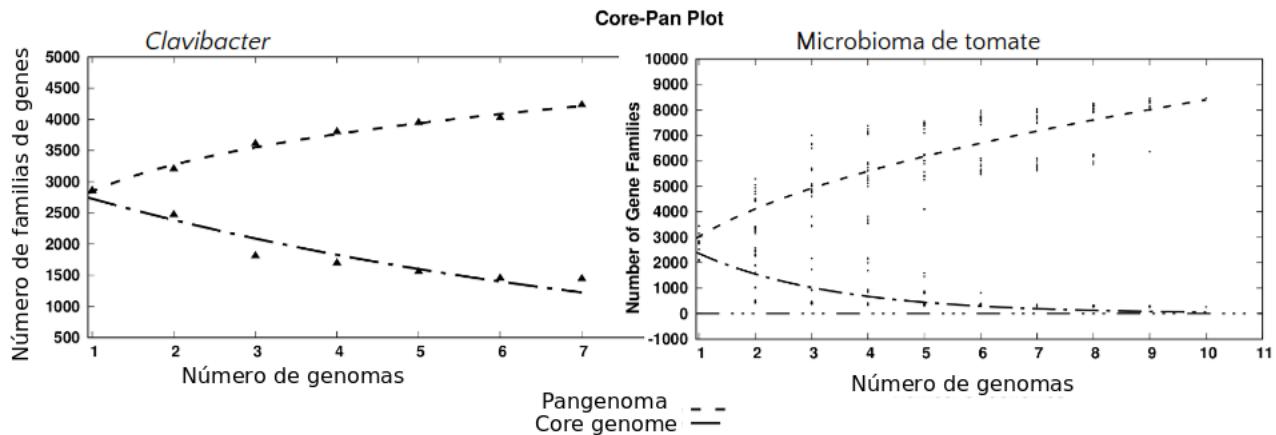


Figure 1.10: adihipihfd

Tomando otros siete clavibacters, utilizando OrthoVenn tenemos la misma observación, el número de familias de genes agregadas al adicionar genomas, es después de siete genomas casi tan grande como su core Figure 1.11.

## 1.5 Relación entre genes marcadores, Orthocores y la promiscuidad enzimática.

Finalmente, al aplicar Orthocore para detectar genes marcadores se vuelve indirectamente reclutamientos al metabolismo especializado, cómo, pues porque dentro de los marcadores hay productos naturales como, PENDIENTE VER RESULTADOS RETRO-EVOMINING la clavidicina (michiganina). Ahí vemos que clvABCDEF participan en metabolismo secundario, las enzimas de metabolismo central de este cluster pueden presentar cierta promiscuidad.

## 1.6 Protocolos para usar Orthocore, myRAST, fastOrtho, Clavigenomics, y BPGA

Anotación genómica con el docker myRAST

Esta es una distribución de myRAST en un contenedor de docker. Para usarla se necesita una cuenta del anotador genómico RAST. el docker myRAST permite hacer

## 1.6. Protocolos para usar Orthocore, myRAST, fastOrtho, Clavigenomics, y BPG5A

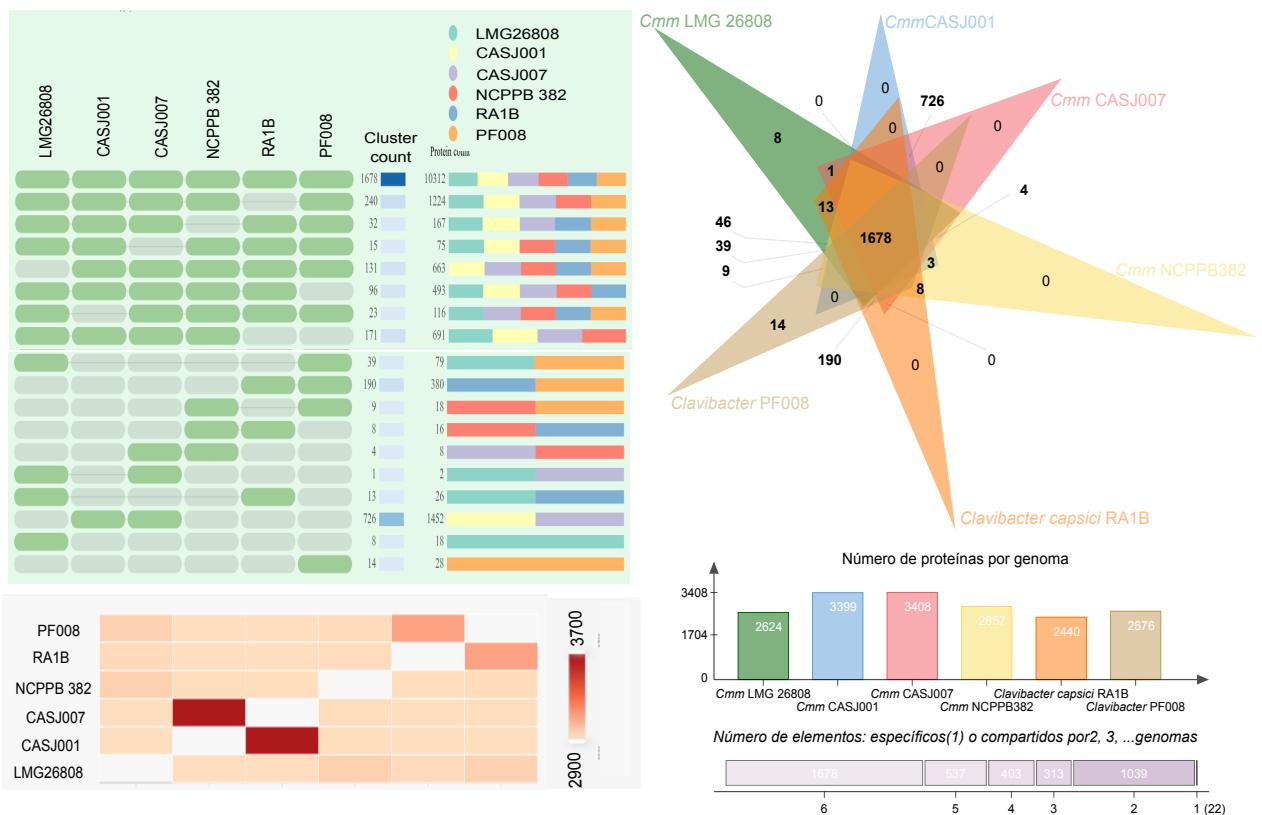


Figure 1.11: Diagrama de Venn del pangenoma de Clavibacter Selectos

anotación genómica y funcional masiva mediante el uso de la terminal en el anotador RAST. Después de anotar los resultados pueden descargarse y procesarse en una terminal

Descargar myrast docker distribution

Una vez con docker instalado en la computadora, se hace pull al docker myrast.

docker pull nselem/myrast

Abrir myRast en la terminal

docker run -i -t -v \$(pwd):/home nselem/myrast /bin/bash

Usar myRast

-Ejemplo subir un archivo fasta

svr\_submit\_RAST\_job -user -passwd -fasta -domain Bacteria -bioname “Organism name” -genetic\_code 11 -gene\_caller rast

-Para bajar un archivo de anotación genómica funcional:

svr\_retrieve\_RAST\_job table\_txt > \$ID.txt

Una lista completa de archivos puede ser procesada usando bash. Por ejemplo para bajar una lista de archivos de RAST se deben guardar los identificadores de RAST en una columna de un archivo, (Rast\_ID en este ejemplo) y usar un while para obtenerlos:

En este caso la variable “line” contendrá el identificador RAST Id, y cada archivo fasta de aminoácidos podrá ser obtenido mediante su identificador de RASTy será guardado en el archivo “\$line.faa”

```
cut -f1 Rast_ID | while read line; do svr_retrieve_RAST_job $line amino_acid > $line.faa ; done
```

Formatos de RAST para descargar archivos

Puedes cambiar el formato table\_txt por el que tú necesites.

Pie de tabla

Table 1.1: Correlation of Inheritance Factors for Parents and Child

Atributo	Descripción
genbank	GenBank (con funciones y enriquecimiento de SEED)
genbank_stripped	Genbank con EC-numbers removidos de las funciones
embl	EMBL (con funciones y enriquecimiento de SEED)
embl_stripped	EMBL con EC-numbers removidos de las funciones
gff3	GFF3
gff3_stripped	GFF3 con EC-numbers removidos de las funciones
gtf	GTF
gtf_stripped	GTF con EC-numbers removidos de las funciones

Atributo	Descripción
rast_tarball	Archivo comprimido (gzipped) con todo el directorio de las anotaciones de RAST sobre el genoma
nucleic_acid	Fasta de DNA de genes
amino_acid	Fasta de DNA de aminoácidos
table_txt	Gene data in tab-separated format
table_xls	Preserve the original gene calls and use RAST

Referencia Table 4.3.

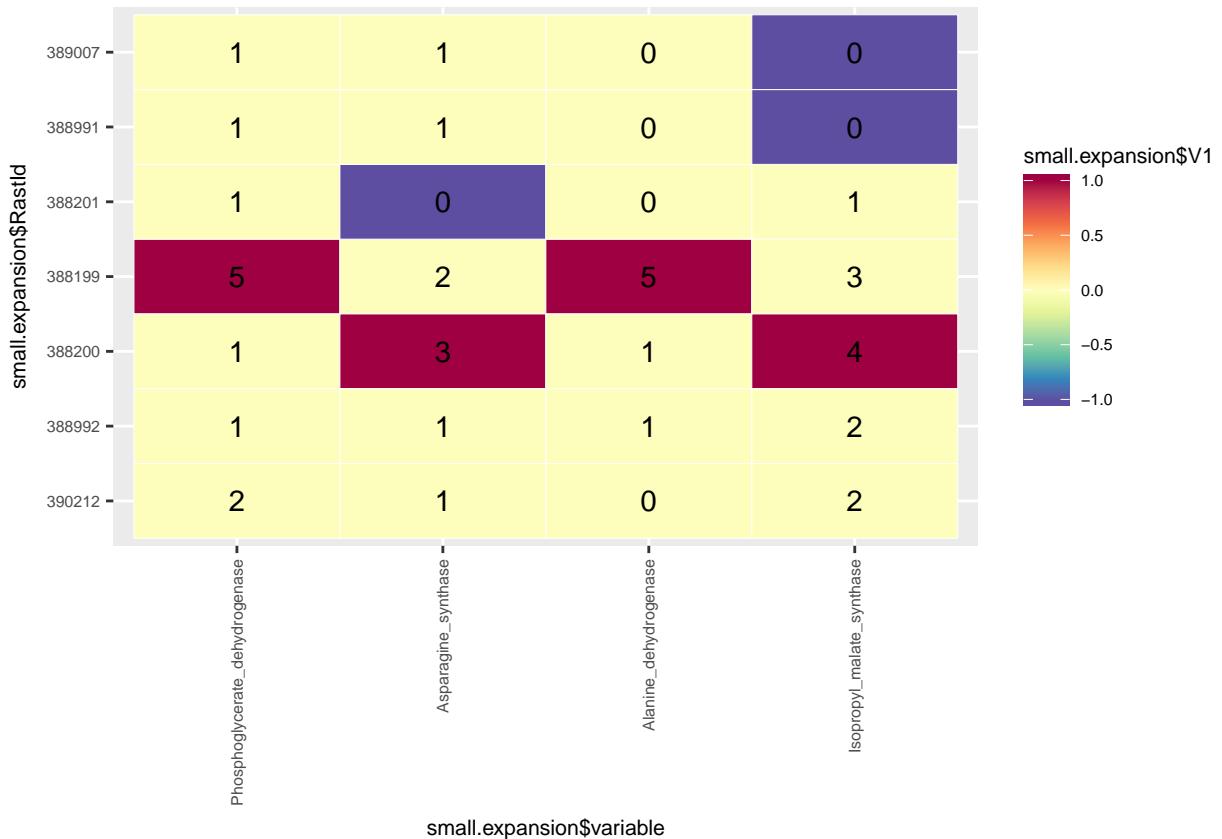
## 1.7 ORthocore

El umbral de e-value de Orthocore es por default 1e-6. Todas las secuencias son alineadas usando MUSCLE v3.8.31 con los parámetros default y curadas utilizando Gblocks con 5 posiciones como longitud mínima del bloque, 10 como máximo número de posiciones contiguas no conservadas y sólo considerando posiciones con gaps menores que el 50% de las secuencias en el alineamiento final. Después de esta curación las secuencias son concatenadas en una matriz final.



# Chapter 2

## EvoMining



```
#####
## Trying to sort the heatplot
## Reading heatplot table and taxa information and saving it on data.frame data st
ArchaeasHeatPlot <- read.table("chapter2/Archaeas/ArchaeasHeatPlot", header=TRUE,
ArchaeasTaxa <- read.table("chapter2/Archaeas/ArchaeasTaxa", header=TRUE, sep="\t"
hm.palette <- colorRampPalette(rev(brewer.pal(11, 'Spectral'))), space='Lab'
## Adding order variable
```

```

ArchaeasHeatPlot$order<-c(1:nrow(ArchaeasHeatPlot))
#sorting RastId it accordig to order variable
ArchaeasHeatPlot$RastId <- with(ArchaeasHeatPlot,reorder(ArchaeasHeatPlot$RastId, ArchaeasHeatPlot$Order))
# Merging heatplot and taxonomy table into one table
HP_Archeas_Taxa<-merge(ArchaeasHeatPlot,ArchaeasTaxa,by.x = "RastId",by.y = "RastId")
# Melting information letting as variables just enzymatic families and copy number
HP_Archeas_Taxa.m<- melt(HP_Archeas_Taxa)

```

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

```

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

HP_Archeas.calcs<- ddply(HP_Archeas_Taxa.m, .(variable),summarize, mean = round(mean(value),3))

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

```

	RastId	Name
1	388199	Haloferax sp ATCC BAA-644ATCC BAA-644 AOLF01
2	388200	Candidatus Methanoperedens nitroreducensANME-2d JMIY01
3	388201	Marine group II euryarchaeote REDSEA-S40_B11N13 LURX01
551	388991	Uncultured Acidilobus sp MG AYMA01
552	388992	Candidate divison MSBL1 archaeon SCGC-AAA259005 LHXV01
567	389007	Uncultured Acidilobus sp OSP8 AYMC01
690	390212	Archaeoglobus fulgidus DSM 8774 DSM 8774 CP006577.1
14017	388199	Haloferax sp ATCC BAA-644ATCC BAA-644 AOLF01

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

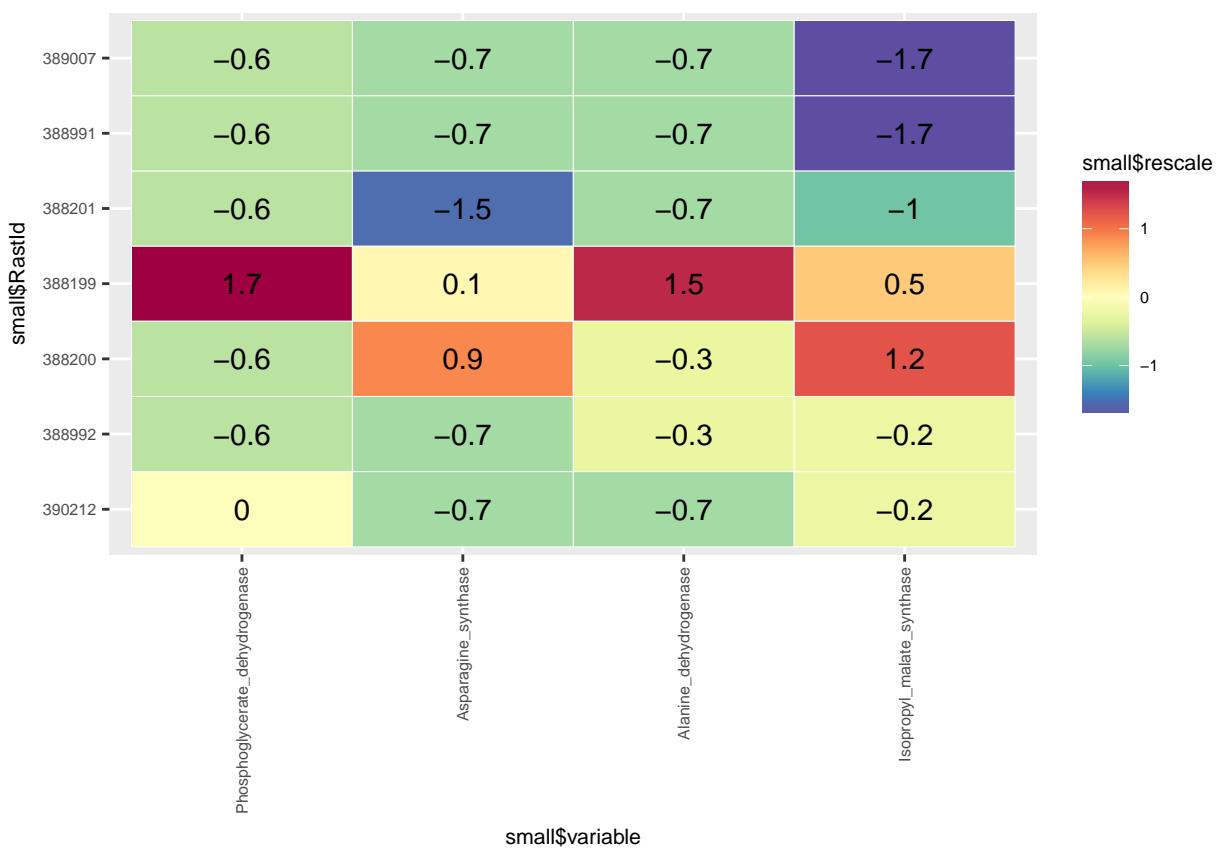
				variable	value	rescale
68879	TACK group Crenarchaeota	Thermoprotei	Acidilobales			
68880	Euryarchaeota Euryarchaeota	unclassified	unclassified			
68895	TACK group Crenarchaeota	Thermoprotei	Acidilobales			
69018	Euryarchaeota Euryarchaeota	Archaeoglobi	Archaeoglobales			
		Family				
1	Haloferacaceae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

```

##Bueno aqui voy
##Heat.expansion<-adply(HP_Archaeas_Taxa.m[!HP_Archaeas_Taxa.m$variable %in% c("Contigs"
#,
#####
#####333
#####

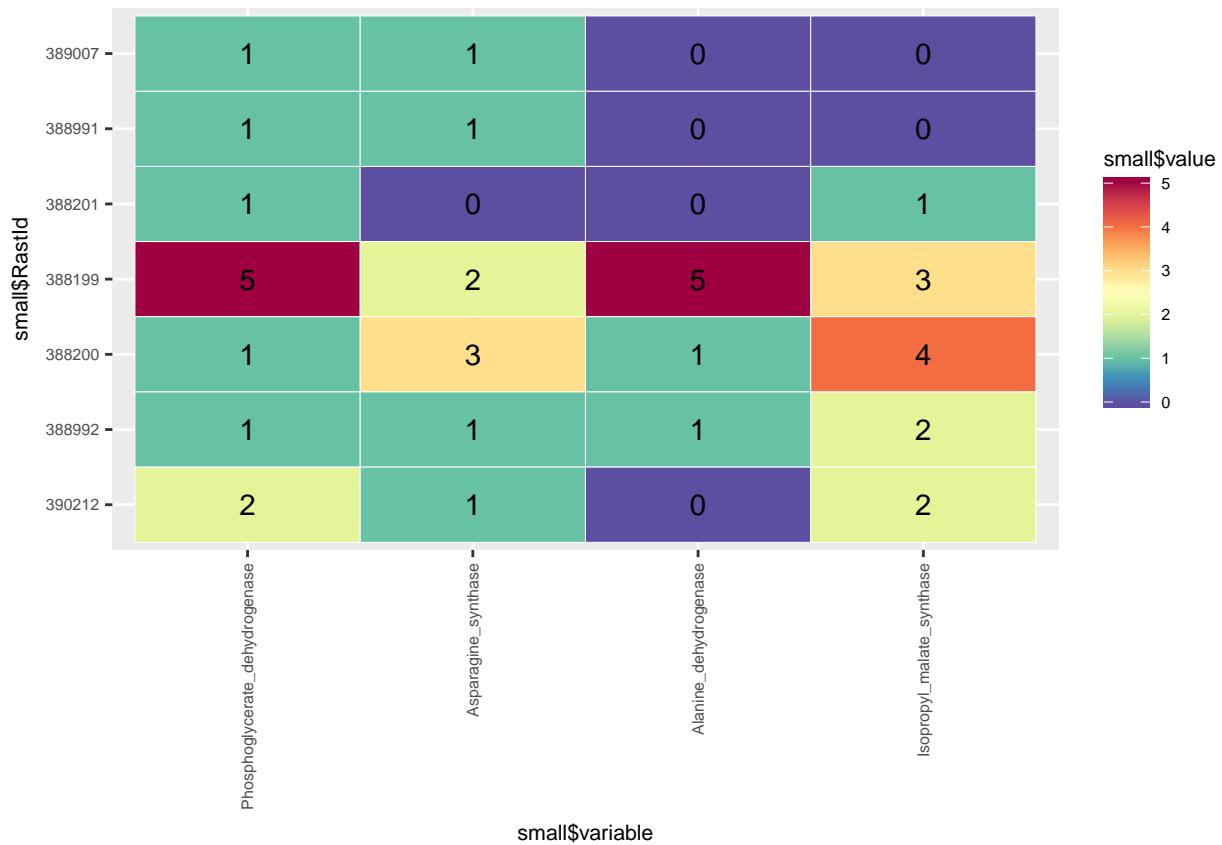
## geom tile with rescale rescala por column
# Graph! with rescale
ggplot(small, aes(small$variable, small$RastId, label=round(small$rescale,digits=1)))+

```



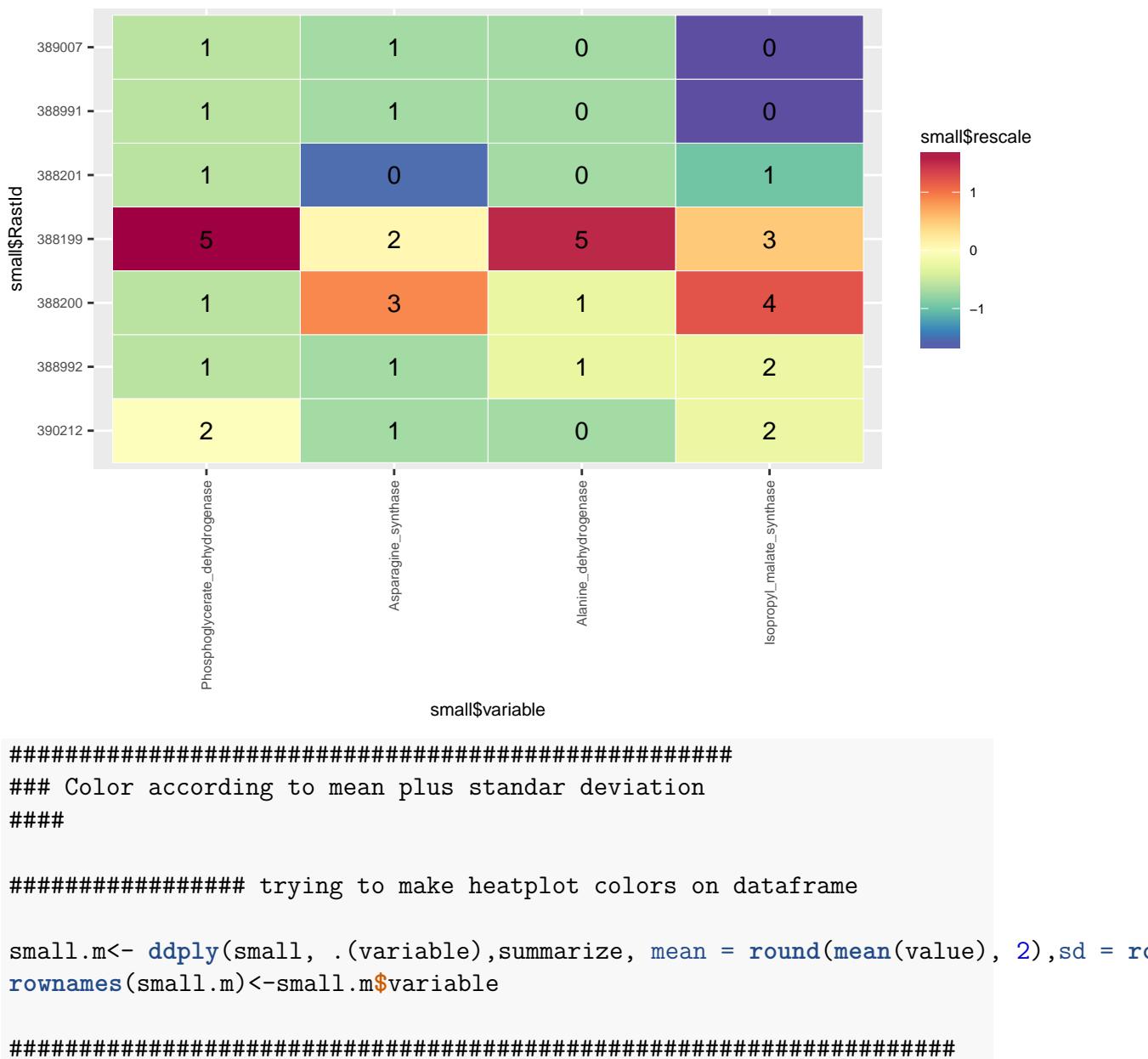
```
## Graph scaled according to whole matrix
```

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



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

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





# Chapter 3

## EvoMining

### 3.1 Introduction

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

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

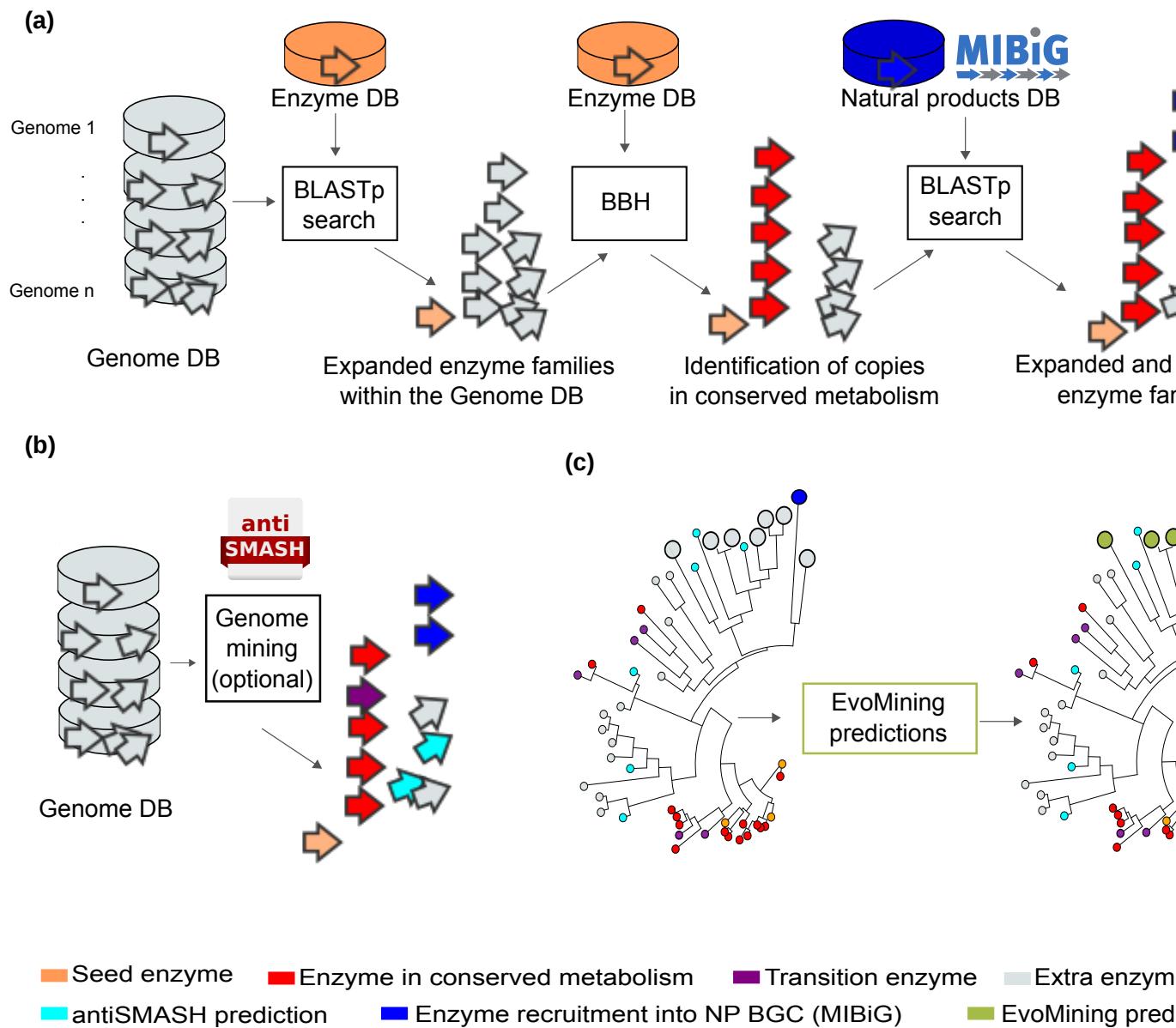


Figure 3.1: EvoMining Algorithm

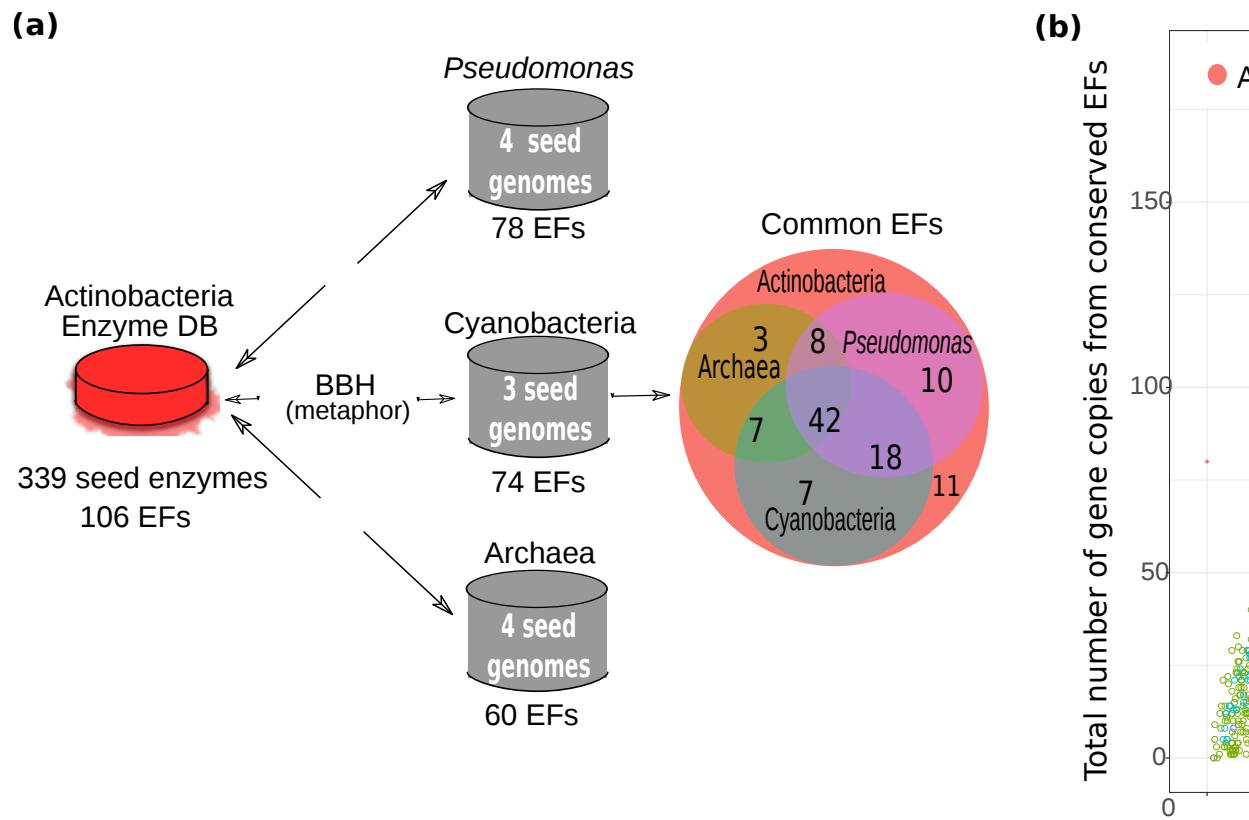


Figure 3.2: Seed genomes

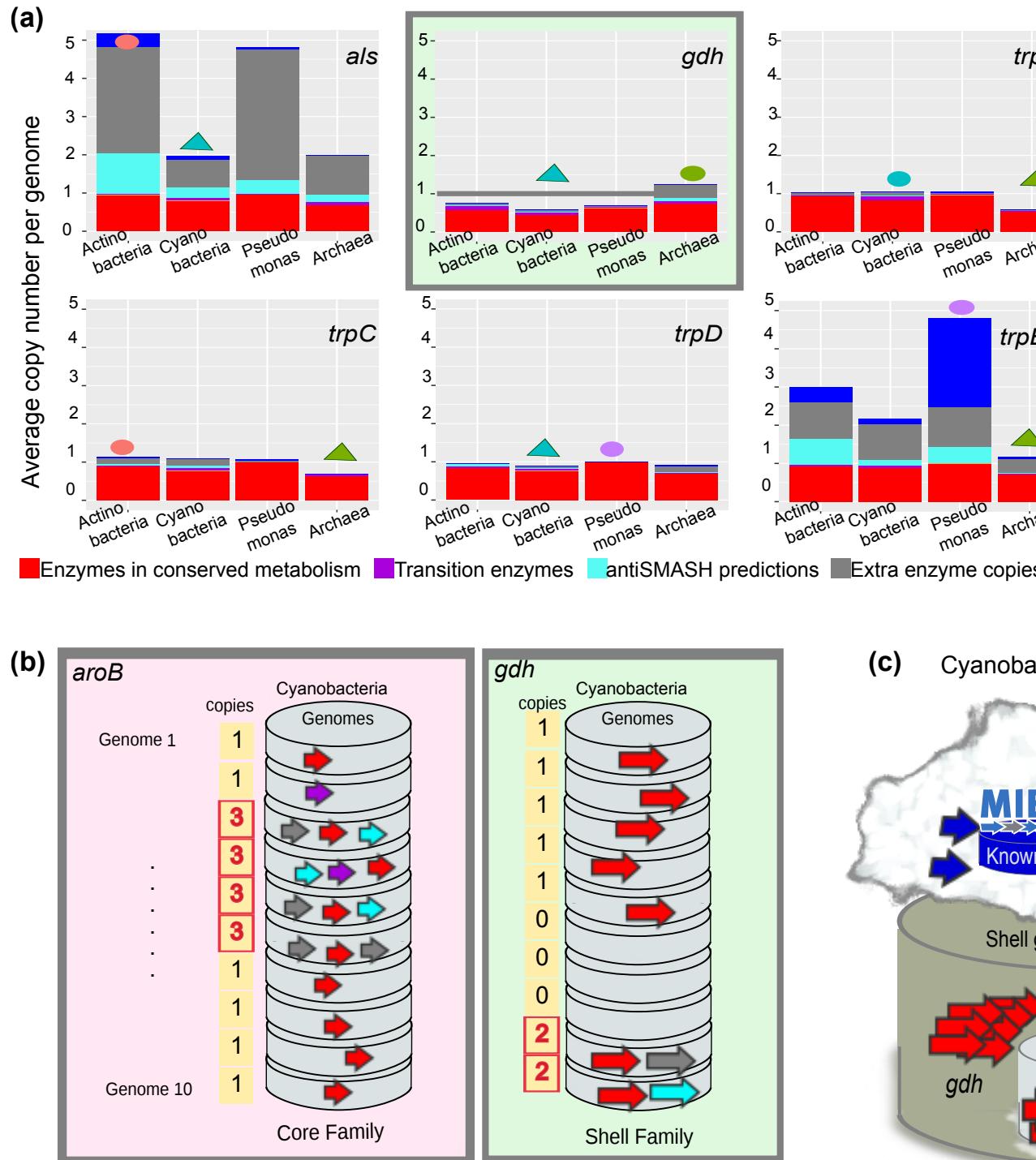
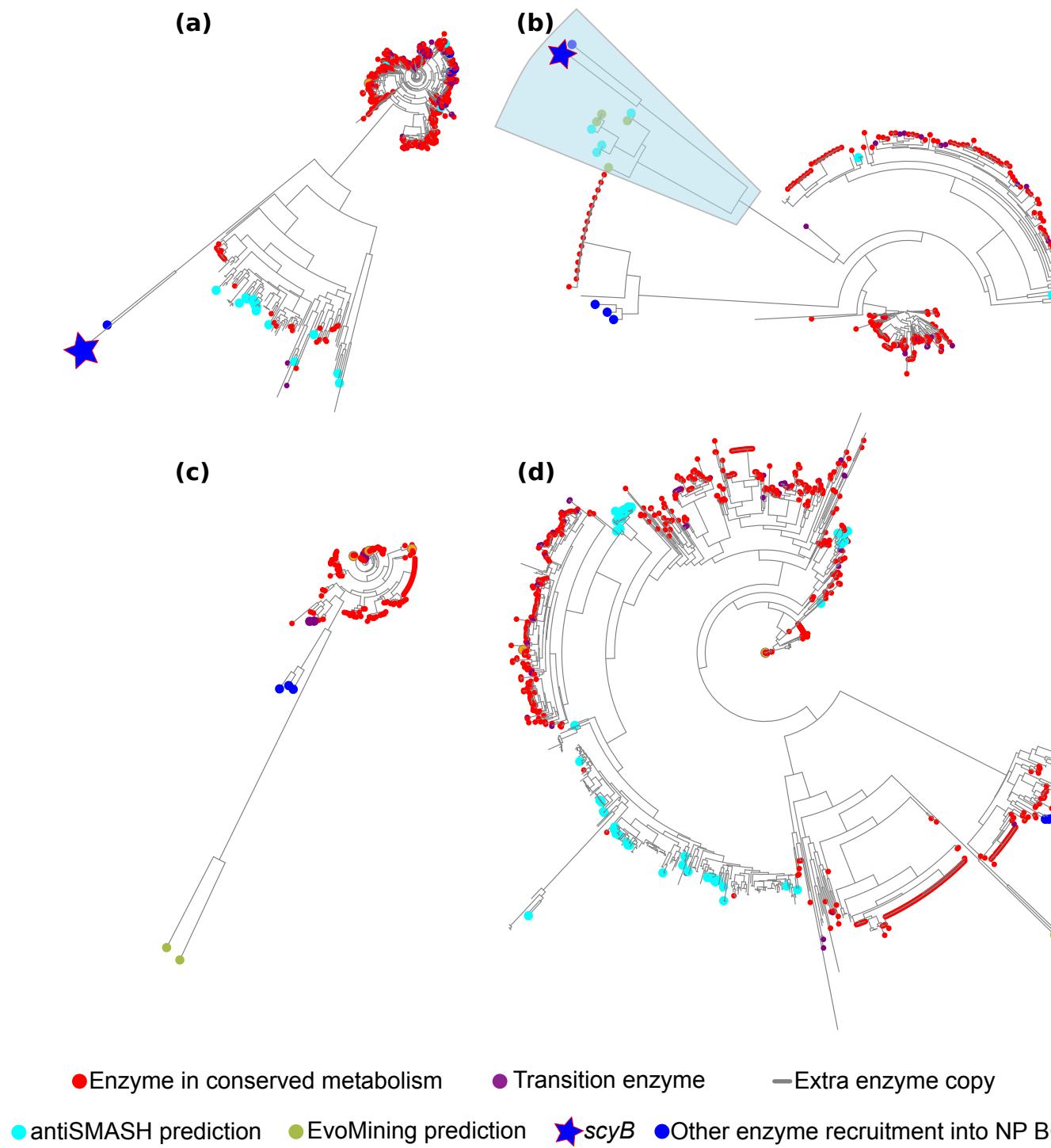


Figure 3.3: Expansion patterns in 42 conserved families



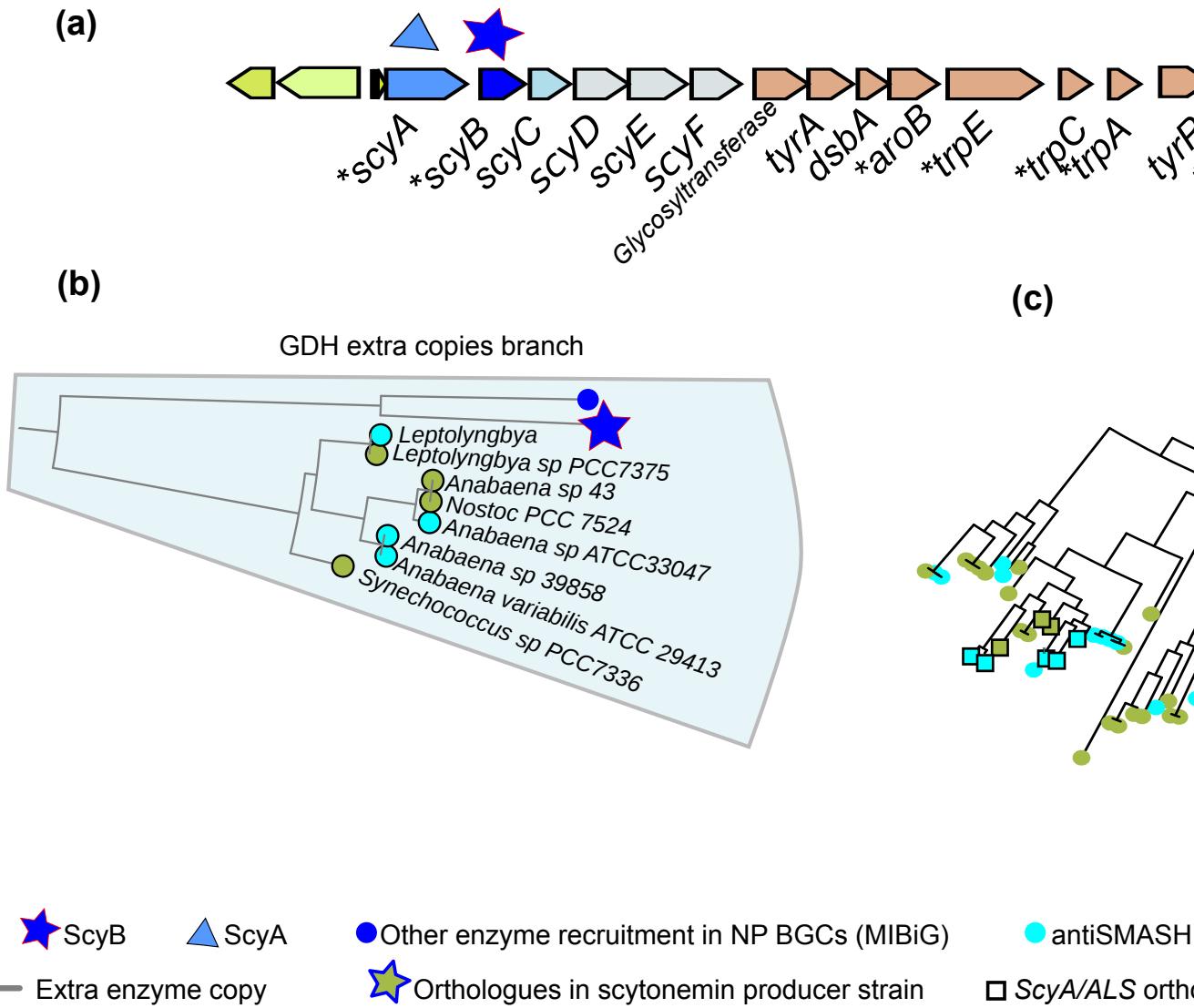
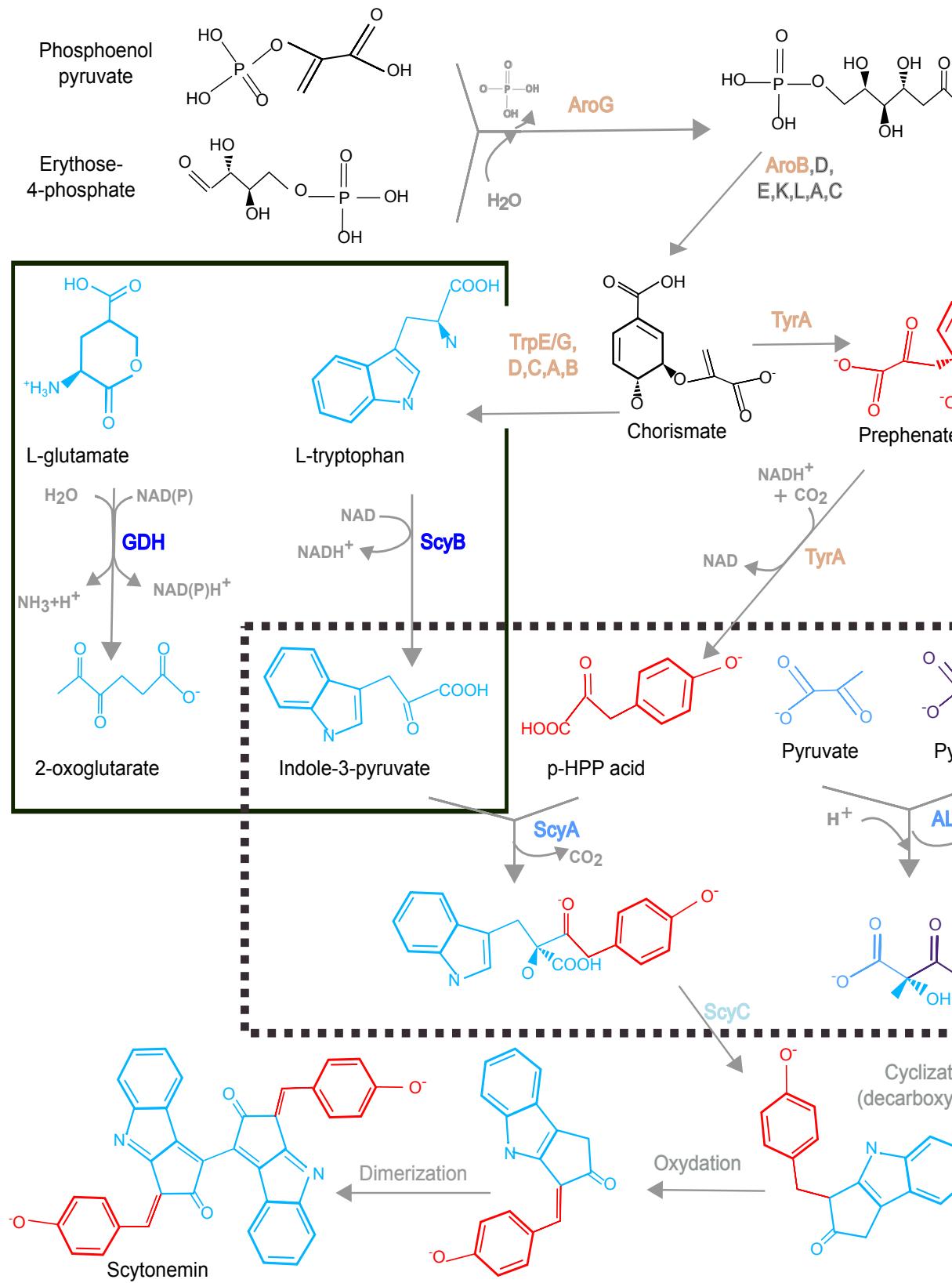


Figure 3.5: EvoMining Algorithm



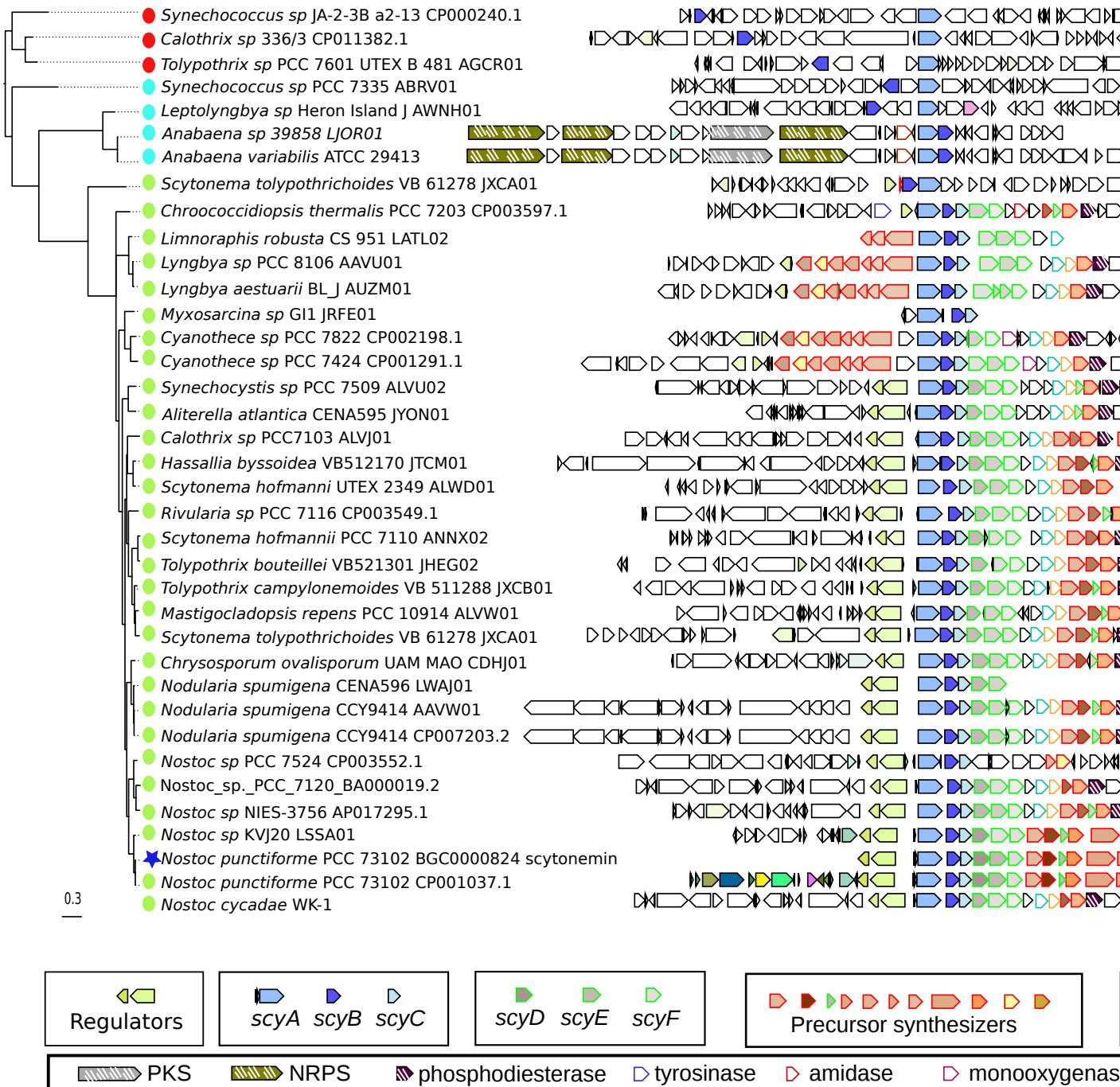


Figure 3.7: EvoMining Algorithm

## 3.2 EvoMining es un método para encontrar BGCs no tradicionales

### 3.3 Gen families expansions on genomes

#### 3.3.1 Pangenomes

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

### 3.4 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 nBGCs, and not interchangeable central database 339 queries for nine pathways, including amino acid biosynthesis, glycolysis, pentose phosphate pathway, and tricarboxylic acids cycle. [52] EvoMining was proved on Actinobacteria Arseno-lipids

### 3.5 Pangenome

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

[76] 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 have a core a shell and a unique genes.

Gene history it's 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 archaea has open pangenome. [34]

HGT doubles mutation rate on prokaryotes. [76] Maybe HGT is an selected feature, if is the case, so could be np production.

Some archaea has open pangenome. [34] Shell trees converge to core trees [117]

## 3.6 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. It's 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 [86] 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 analyze 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 concern 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 [143]. While R-markdown allows to write and run R code and interpolate text paragraph to explain scripts and analysis.

## 3.7 EvoMining Databases

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

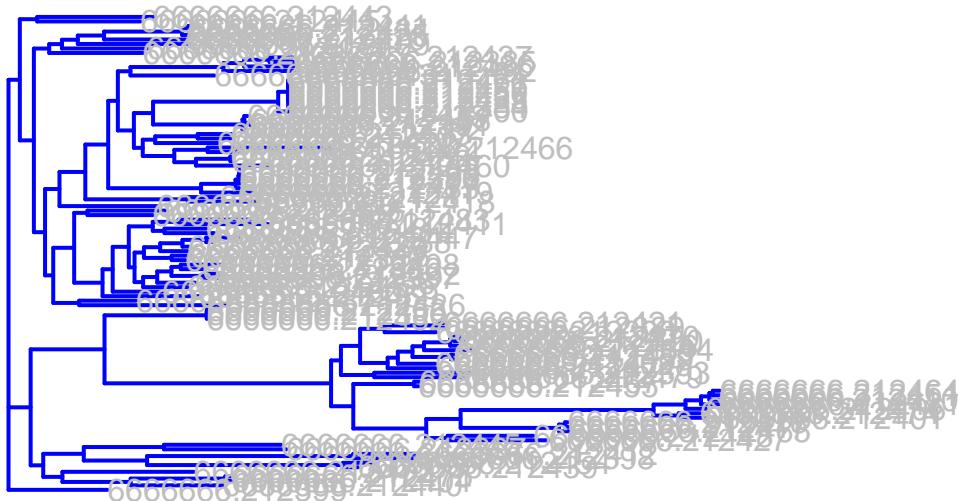
1. Genomes DB
2. Natural Products DB
3. Central Pathways DB

*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 [144]

\* 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/]
```

## 3.8 Data Bases

### 3.8.1 Central pathways

Central database were chosen by BBH from

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

Table 3.1: 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	
Methanosaerica 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
Cyanophage sp. ATCC 51142 CP000806.1	6666666.212444	Cyanobacteria	Oscillatoriophyta
Synechococcus sp. PCC 7002 CP000951.1	6666666.212477	Cyanobacteria	Synechococcus
Arthrospira platensis C1	6666666.189647	Cyanobacteria	Cyanobacteria

### 3.8.2 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 organims were

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

Table 3.2: WC\_Organisms

	Rast.Id	Database
Arthrospira platensis NIES-39 AP011615.1	6666666.21	Cyanos

	Rast.Id	Database
Synechococcus sp. PCC 7002	6666666.21	Cyanos
Cyanothece sp. ATCC 51142	6666666.21	Cyanos
Methanosaerina 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

### 3.8.3 AntisMASH optional DB

AntiSMASH is [???,???

### 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

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

## 3.9 Argonne

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

cs close strain
wc whats chain

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

in Seqs (not mine)
cat
6666666.103569 6666666.112815 6666666.112823 6666666.112833 6666666.112841
6666666.112849 6666666.112857 > /home/nse/Concat_Full
to find paralogous sets
svr_representative_sequences -b -f Id_Clust -s 0.5 < Concat_Full > TempFull&
perl -p -i -e 's///' readable.tree to clean the tree
To find contexts o pegs of paralogous sets

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

fig|6666666.112839.peg.26

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

cluster\_objects < elements\_to\_cluster > ClusteFile

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

write a file similiar but with the family number

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

compare each peg on this file from the same family

Write the conextions file

peg1 peg2  
peg1 peg3  
peg2 peg3

cluster this file and score the cluster

Define

1. a "function set" is generated by the what's changed directory as a "family"
  2. a "paralog set" is a set of function sets in which paralogous members span the sets
  3. a PEG is in a paralog set if it is in one ofthe function sets that make up the
  4. a "context" of a PEG is the set of close pegs
- 4.1 First cluster operation would give us: context sets (CS)
5. a "context set" is a set of PEGs with "similar contexts"
- 5.1 second clustering operation would give us:cluster (C1)
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}

C1={[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
```

### 3.9.1 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].

## 3.10 Recomendaciones de Luis

Para evoMining

Probar distintos métodos de filogenia y después hacer la coloración.

maximum likelihood, Protest phyml

Atracción de ramas largas.

raxml

trim all vs Gblocks (Tony Galvadon)

Comparar dos árboles

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

Robinson and foulds

Joe Felsenstein

Phylip

2. dist tree

quarter descomposition

peter gogarten fendou Mao

Sets de experimentos.

Para el experimento de los streptomyces con ruta centrales el core, analizar el problema

de dominios múltiples.

Dominios

Nan Song, Dannie durand

Después del blast

Para obtener

Pablo Vinuesa: Get Homologues

Burkhordelias y su toxina (Preguntar a Beto)

Cianobacterias y la ruta de fijación de nitrógeno.

Servidor Viernes a las 12:00

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

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

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

The core genome on a taxonomical group is the set of coding sequences that are shared between all group members, this definition may be adapted to the cluster core by exploring a set of gene clusters instead of a set of genomes. The cluster core attempts to identify a set of functions conserved on a particular BGC variations. A report about gene function using RAST technology will be provided whenever a cluster core exists and core sequences will be concatenated to construct a phylogenetic tree and sort variation clusters accordingly.

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

Finally, in order to provide an easy to install distribution, CORASON was packaged on docker containerization platform. Software dependencies such as BLAST 2.2.30,

muscle3.8.3, GBlocksLinux64\_0.91b, quicktree, newick-utils-1.6, and CORASON code were wrapped together on CORASON docker container. Tutorial and software are available at nselem/github.

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

In conclusion CORASON is an easy to install comparative genomic visual tool on a customizable genome database that allows users to 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.

### 3.13 Tree methods (from antiSMASH textual quotation)

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

# Chapter 4

## EvoMining Results

### 4.1 Archaea

During the decade between 1970 and 1980, Archaea was recognized as new life domain, a kingdom different from Bacteria and Eucarya in an exciting first great application of 16S phylogeny[124]. 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[108].

Natural products biosynthetic gene clusters search is actually performed using either *high-confidence/low-novelty or low-confidence/high-novelty* bioinformatic approaches [44]. 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 [52]. 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 [127] 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 [127]. 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 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

## 4.2 Tables

Table 4.1: 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.

#### 4.2.1 Expansions BoxPlot by metabolic family

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

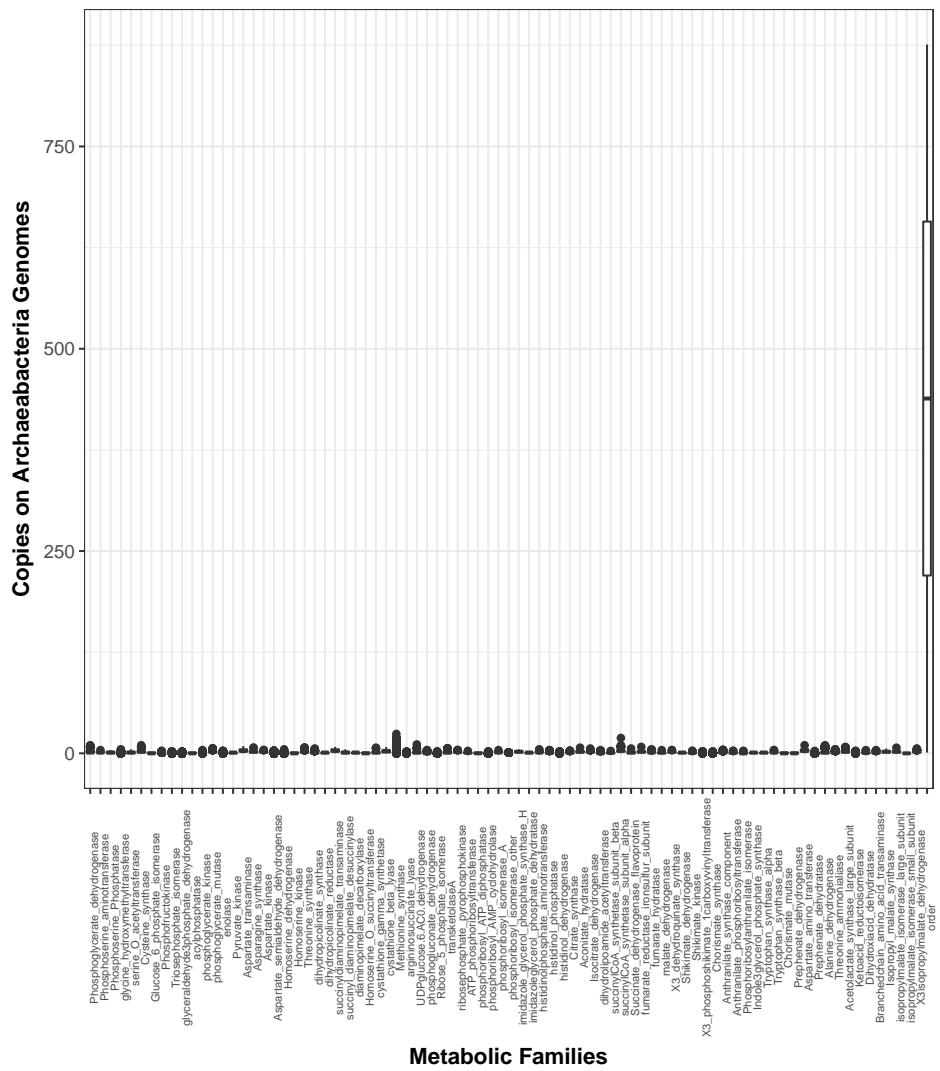


Figure 4.1: Expansions Boxplot

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

#### 4.2.2 Expansions BoxPlot by metabolic family by phylum

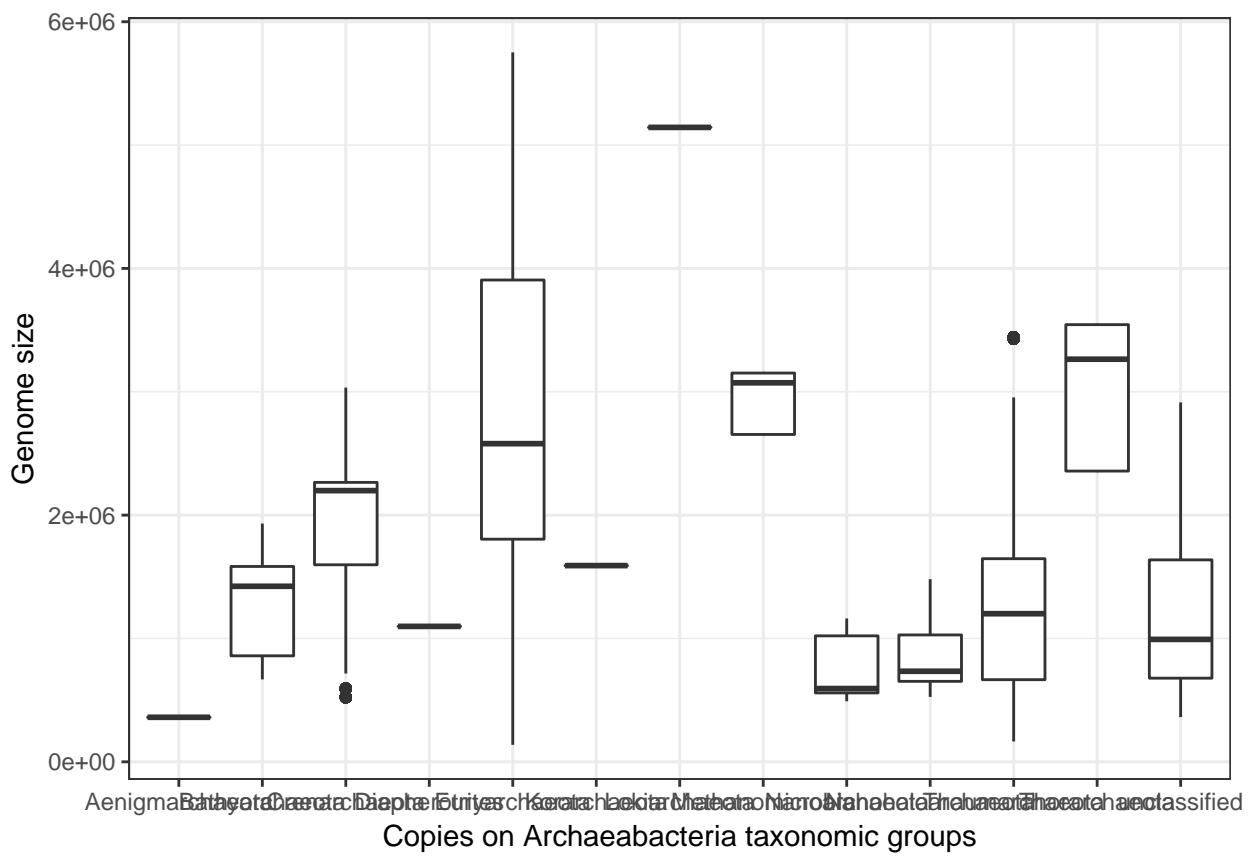
```
#+ geom_jitter()
#aes(fill = factor(vs))

ArchaeasTotalBP.m<-merge(ArchaeasHeatPlot,ArchaeasTaxa,by.x="RastId",by.y="RastId") ## w
##melt in r
ArchaeasHeatPlotBP.m <- melt(ArchaeasTotalBP.m,id =c("RastId","Name","SuperPhylum","Phyl
ArchaeasHeatPlotBP.m<-subset(ArchaeasHeatPlotBP.m,variable!="TOTAL") ## works as expecte
ArchaeasHeatPlotBP.m<-subset(ArchaeasHeatPlotBP.m,variable!="TOTAL") ## works as expecte

## 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
```

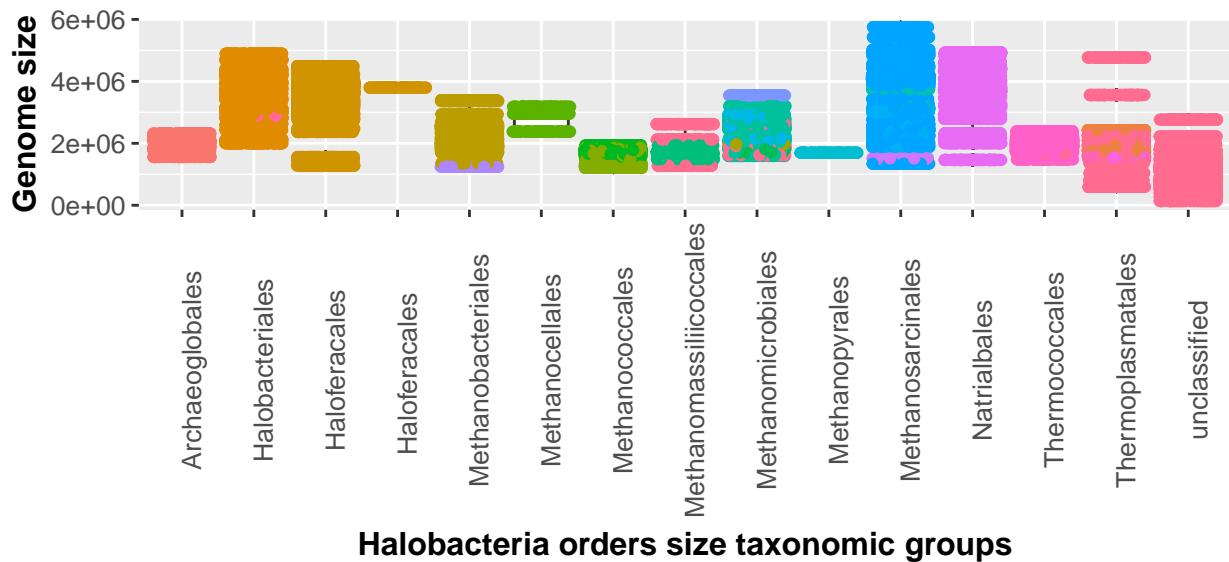


```

## 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()

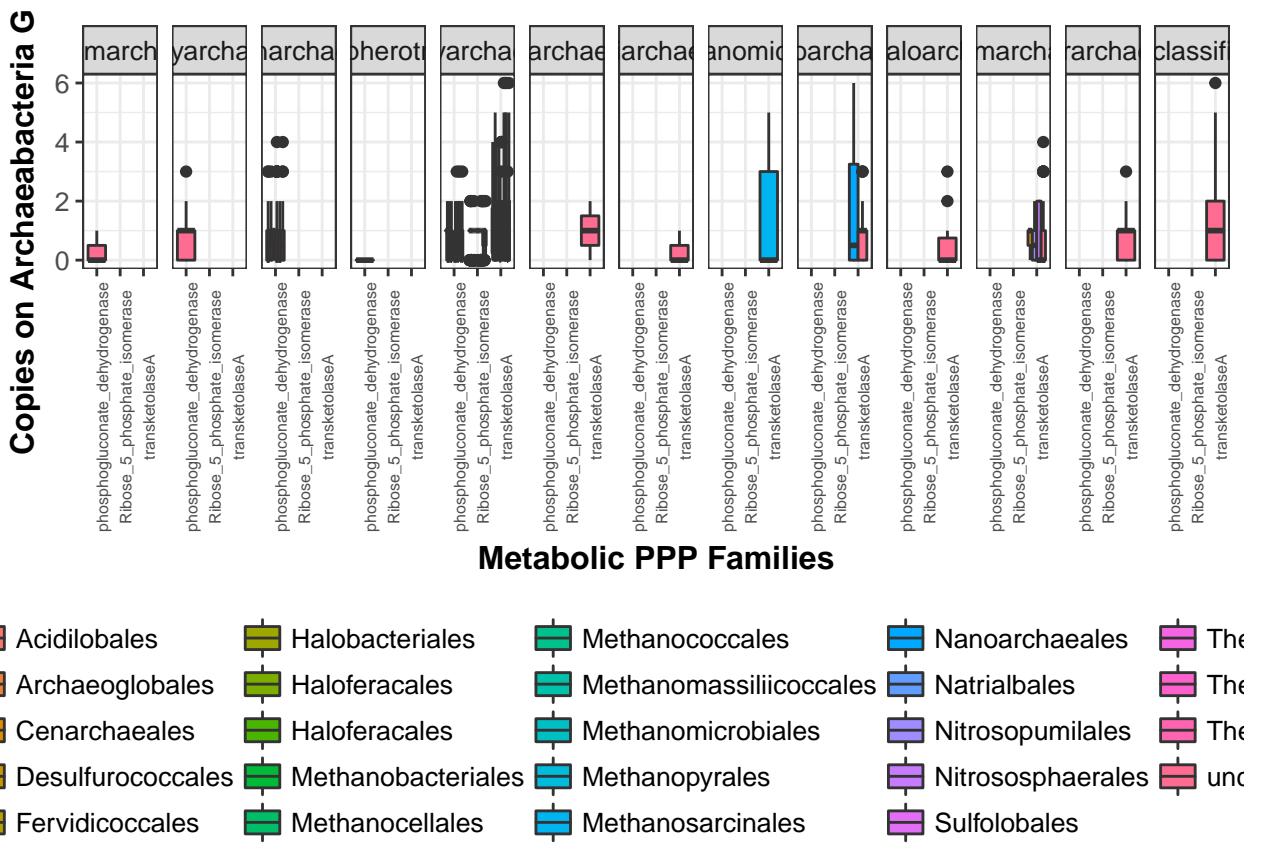
```



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

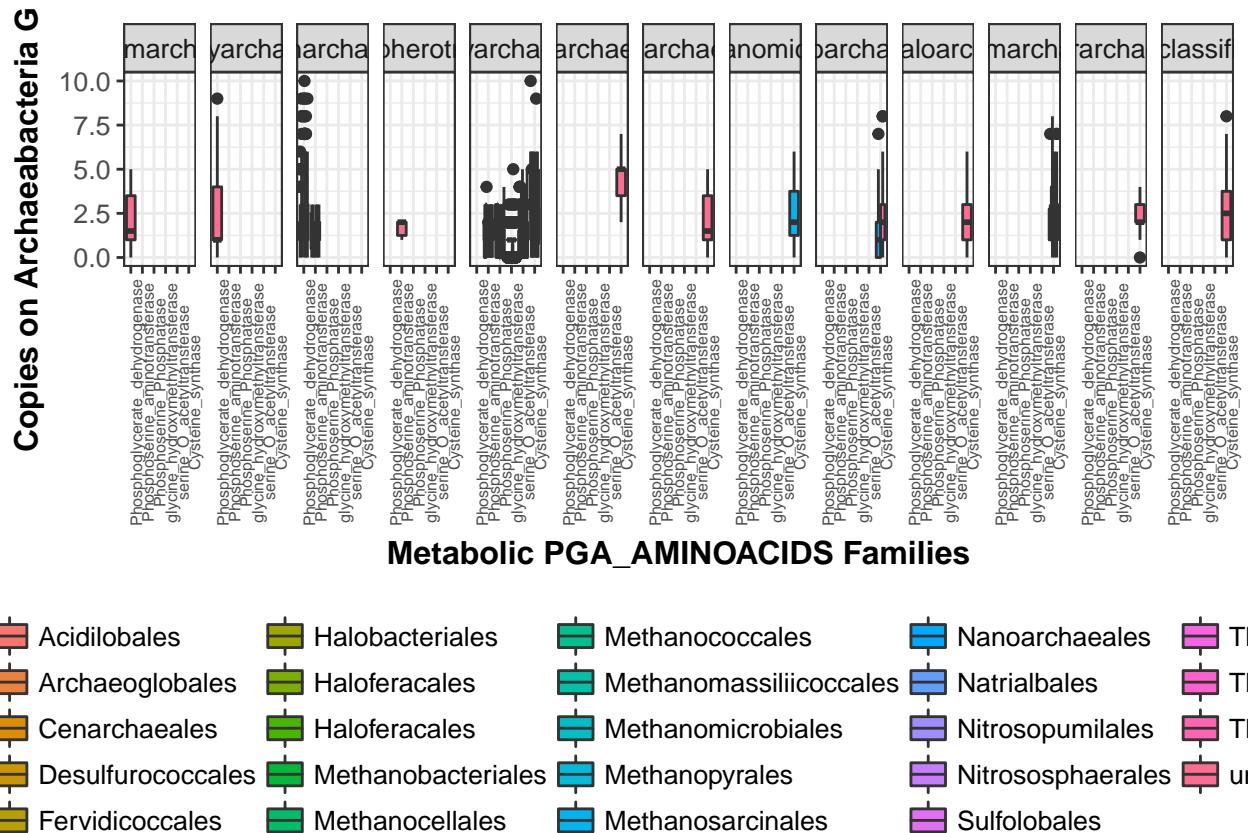
#geom_jitter(aes(color=ArchaeaHeatPlotBP.m$Phylum))# + facet_grid(. ~ Phylum)+theme

## 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 Pathway", y="Percentage")+
  geom_bar(stat="identity")+
  theme_minimal()+
  theme(panel.grid.major.x=element_line(), panel.grid.major.y=element_line())
  
```



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

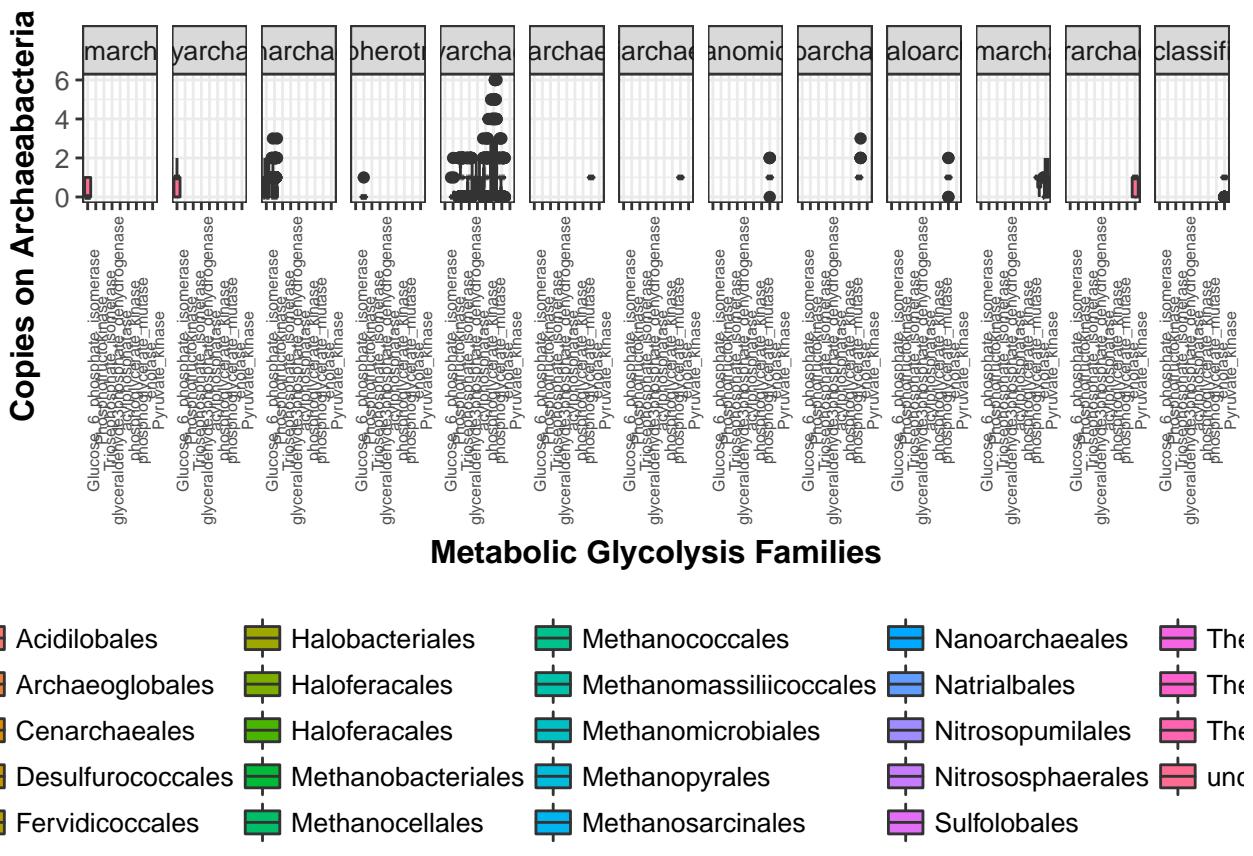
MetFam=subset(ArchaeaCentral,Pathway=="3PGA_AMINOACIDS")
MetFam_BP.m=ArchaeaHeatPlotBP.m[ArchaeaHeatPlotBP.m$variable %in% MetFam$Enzyme,
ggplot(MetFam_BP.m, aes(x=MetFam_BP.m$variable, y=MetFam_BP.m$value, fill=Order)) +
```



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

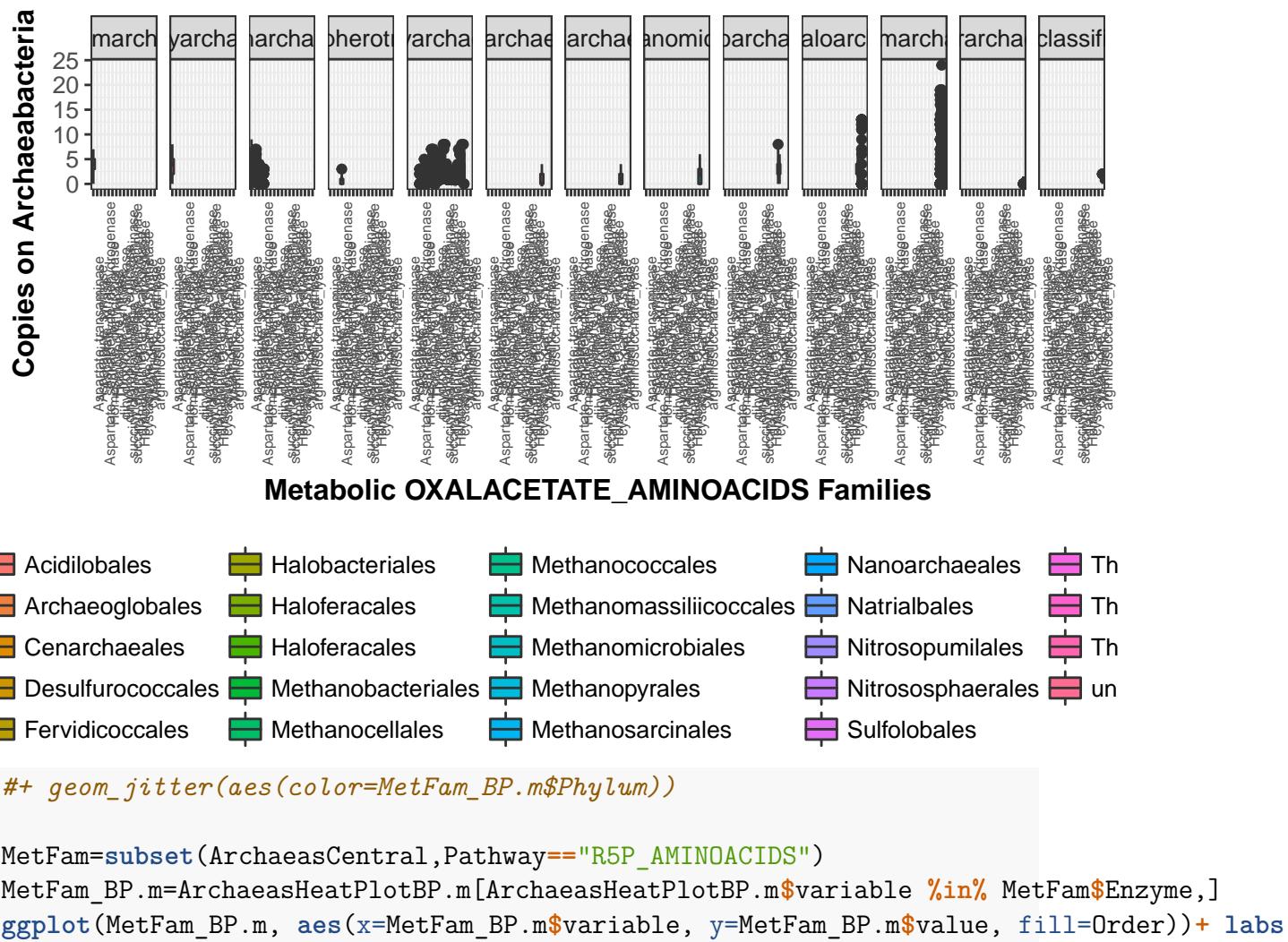
```
MetFam=subset(ArchaeasCentral,Pathway=="Glycolysis")
```

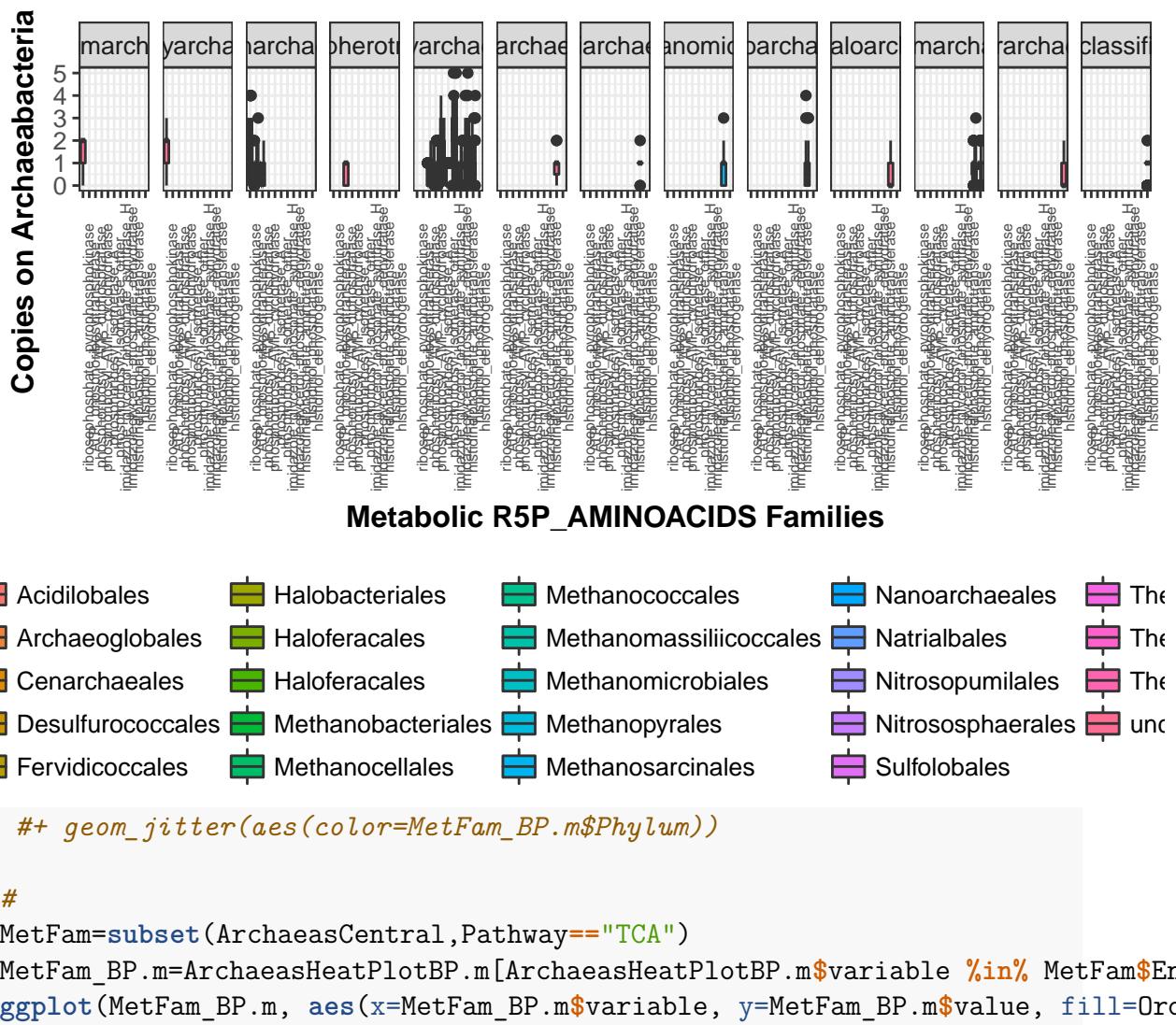
```
MetFam_BP.m=ArchaeasHeatPlotBP.m[ArchaeasHeatPlotBP.m$variable %in% MetFam$Enzyme,]
ggplot(MetFam_BP.m, aes(x=MetFam_BP.m$variable, y=MetFam_BP.m$value, fill=Order))+ labs
```

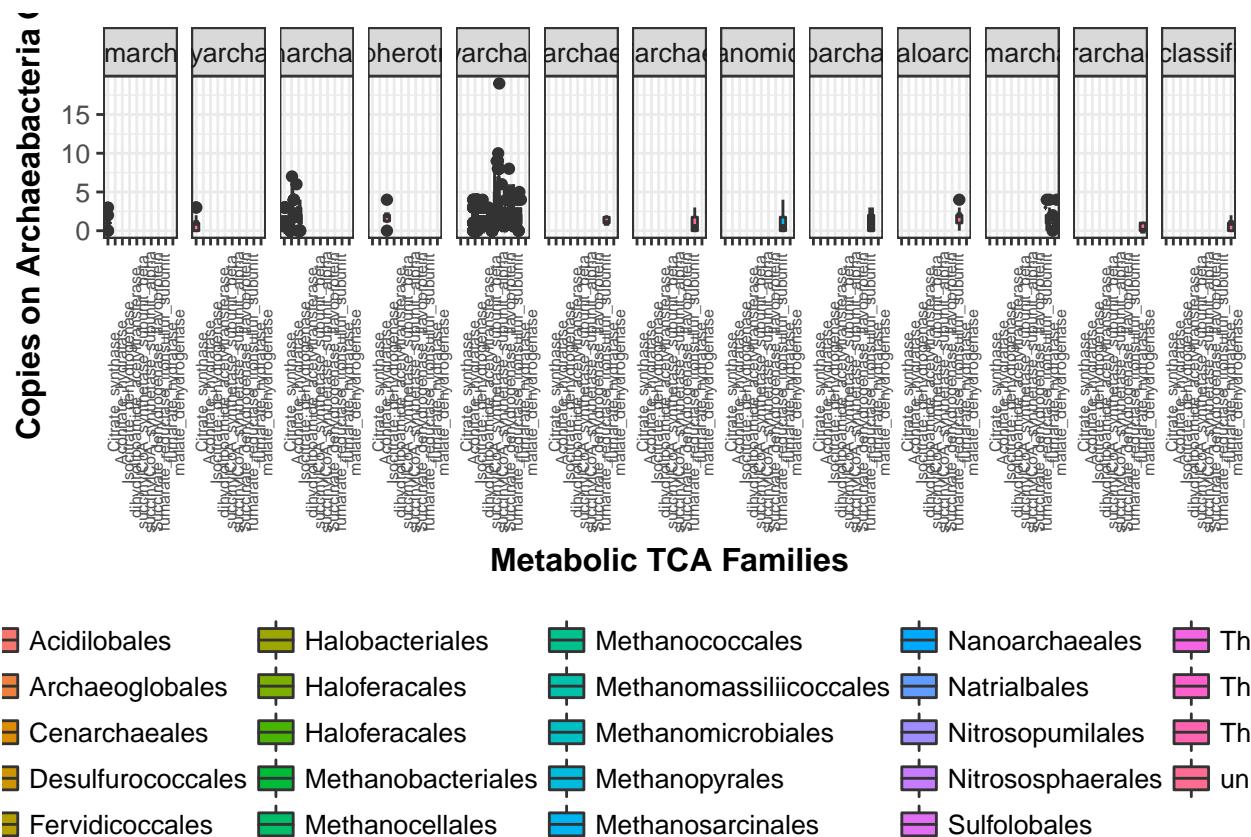


```
#+ geom_jitter(aes(color=MetFam_BP.m$Phylum))

MetFam=subset(ArchaeaCentral,Pathway=="OXALACETATE_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)) +
```

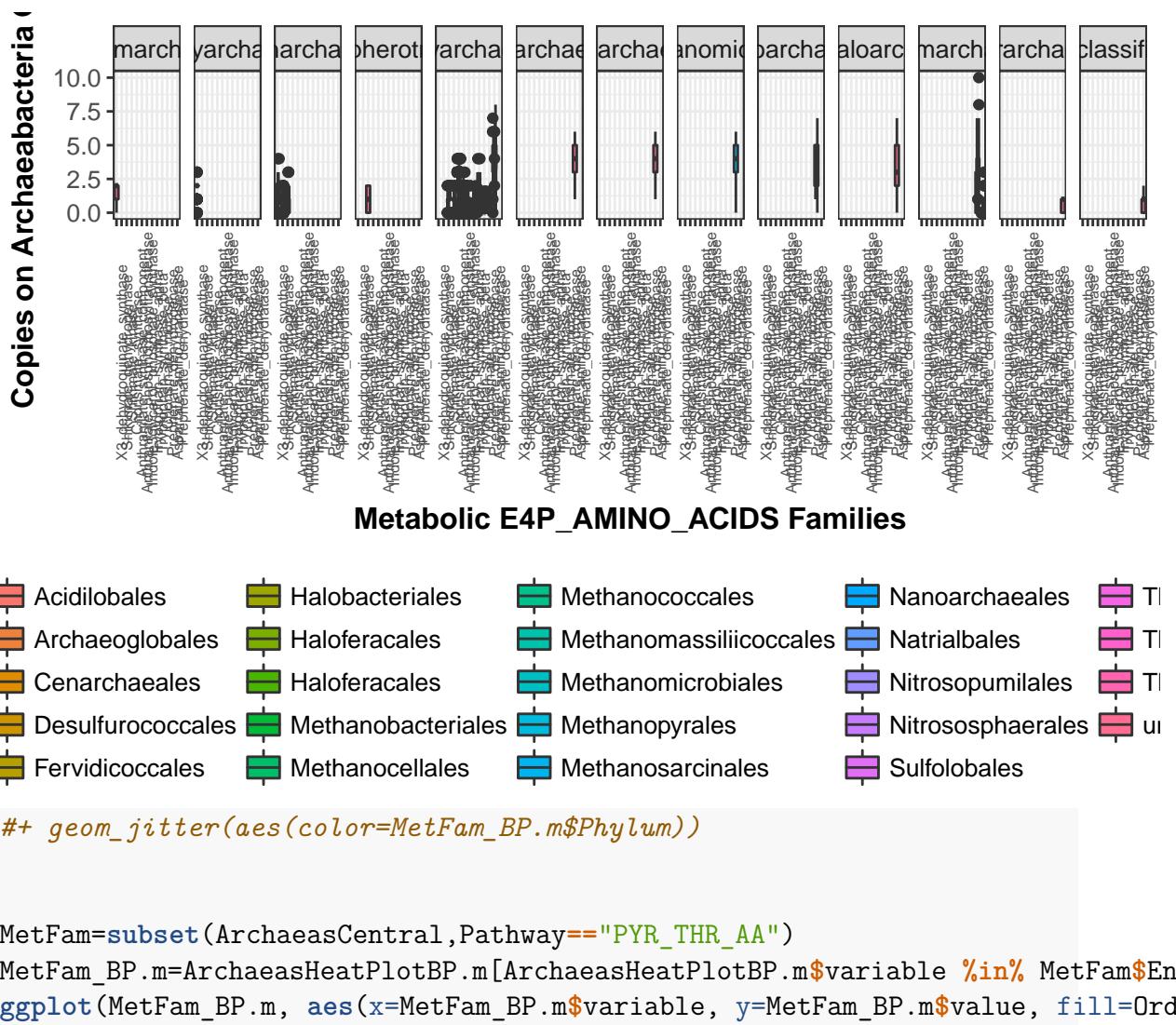


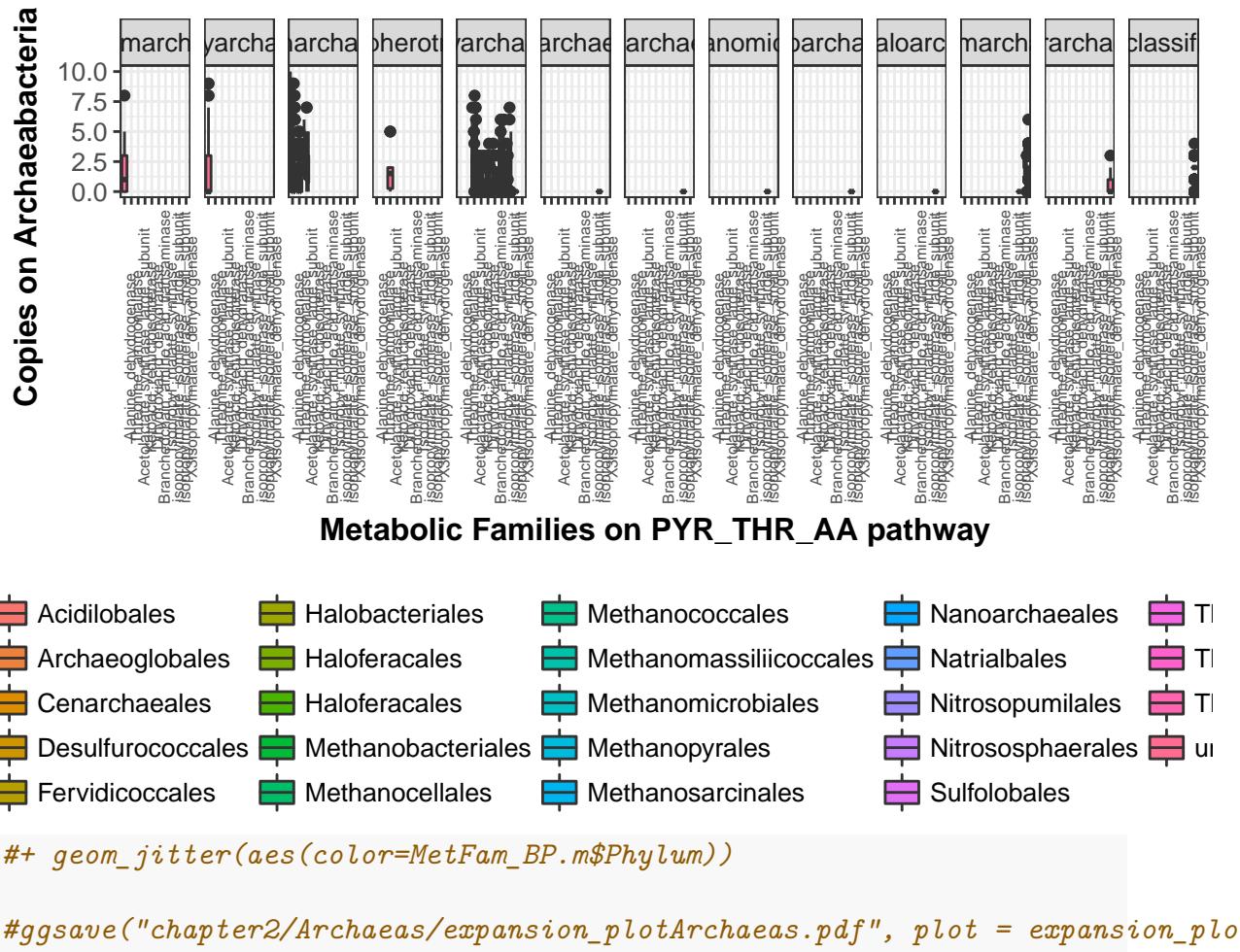




```
#+ geom_jitter(aes(color=MetFam_BP.m$Phylum))

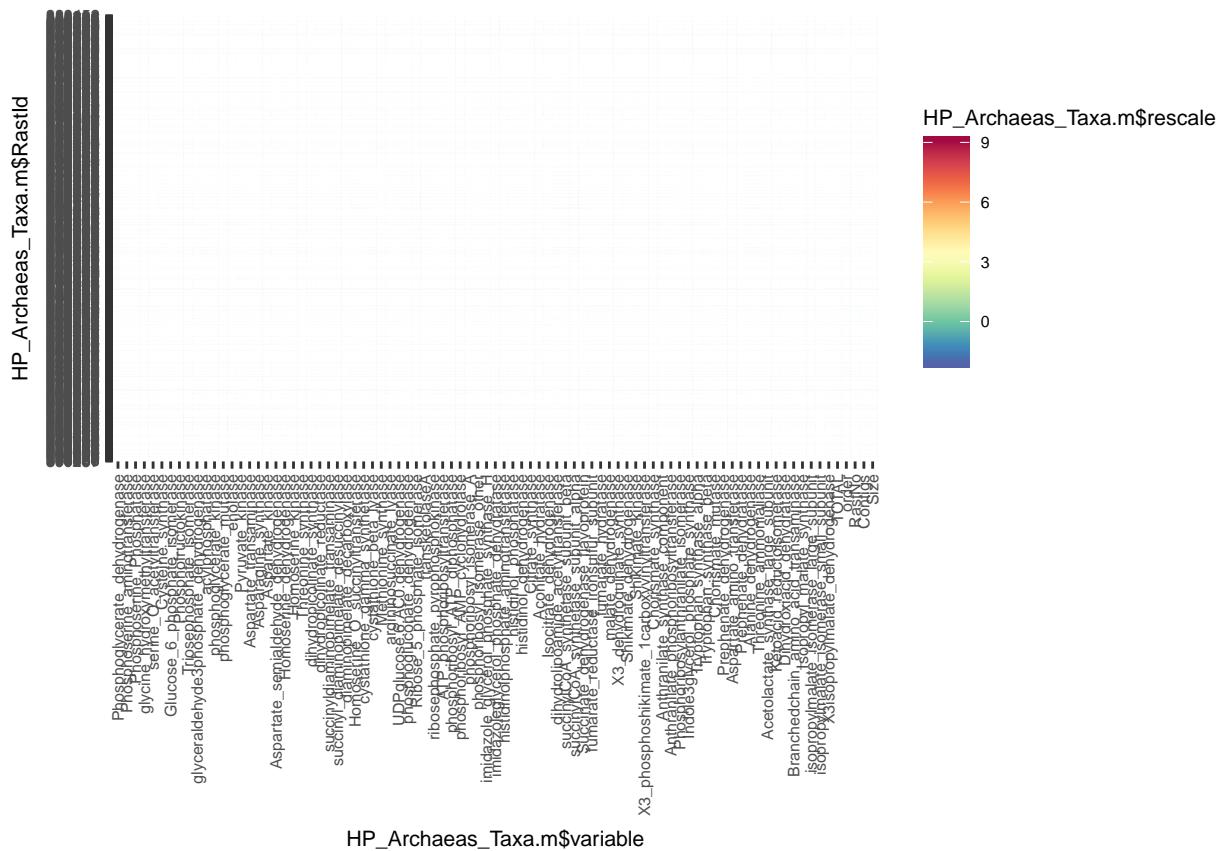
MetFam=subset(ArchaeaCentral,Pathway=="E4P_AMINO_ACIDS")
MetFam_BP.m=ArchaeaHeatPlotBP.m[ArchaeaHeatPlotBP.m$variable %in% MetFam$Enzyme,]
ggplot(MetFam_BP.m, aes(x=MetFam_BP.m$variable, y=MetFam_BP.m$value, fill=Order)) + labs
```





### 4.3 Central pathway expansions

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



Here is a reference to the HeatPlot: Figure 4.2.

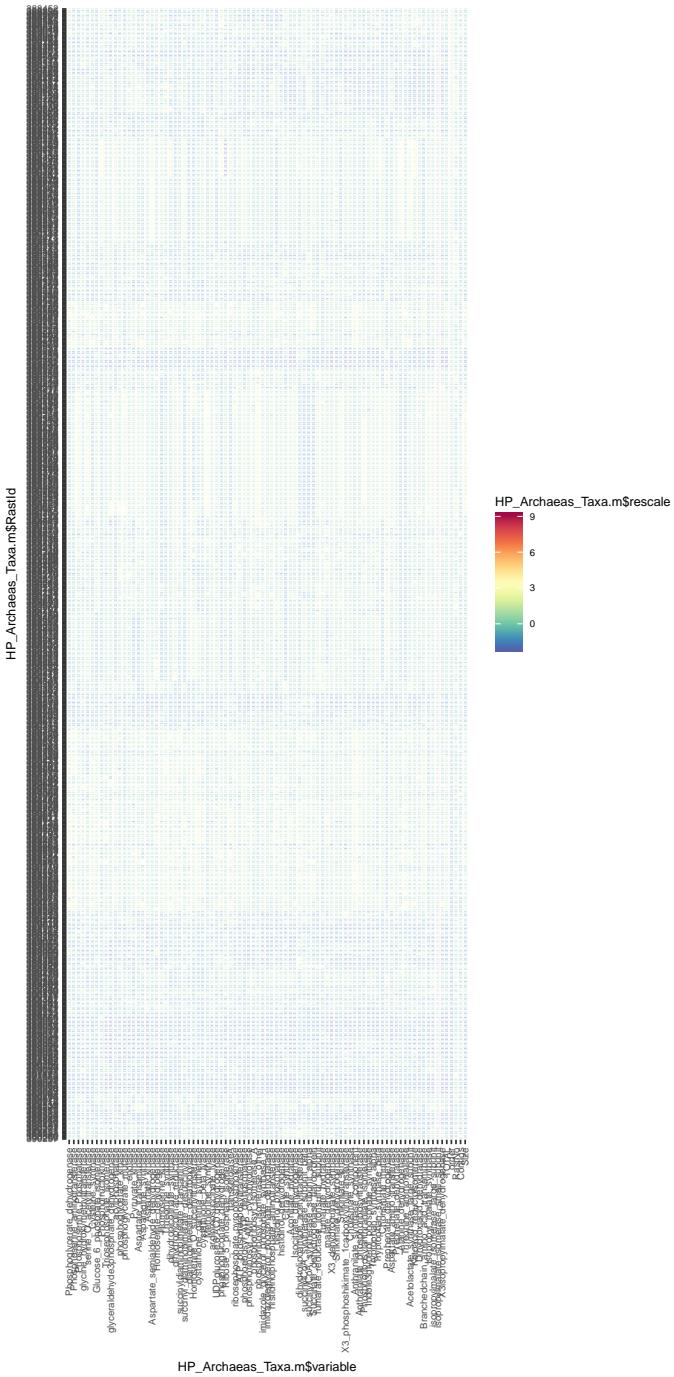


Figure 4.2: Archaeas Heatplot

## 4.4 Genome Size correlations

### 4.4.1 Correlation between genome size and AntiSMASH products

Genome size vs Total antismash cluster coloured by order

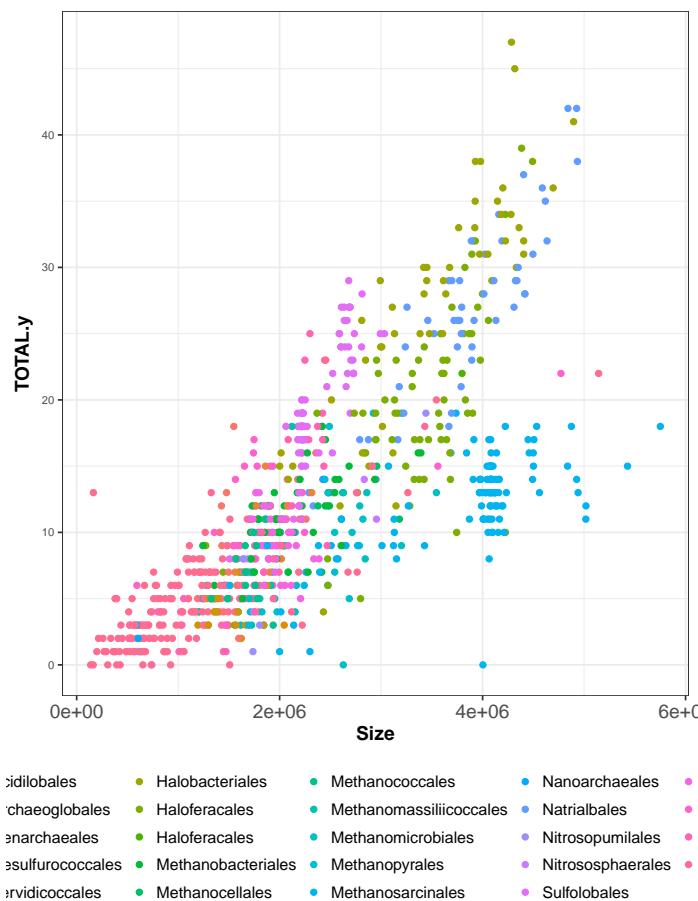


Figure 4.3: Correlation between Archaea genome size and antismash Natural products detection colored by Order

Here is a reference to Genome size vs Total antismash cluster: Figure 4.3.

Genome size vs Total antismash cluster detected splitted by order

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

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

Genome size vs Total central pathway expansion coloured by order

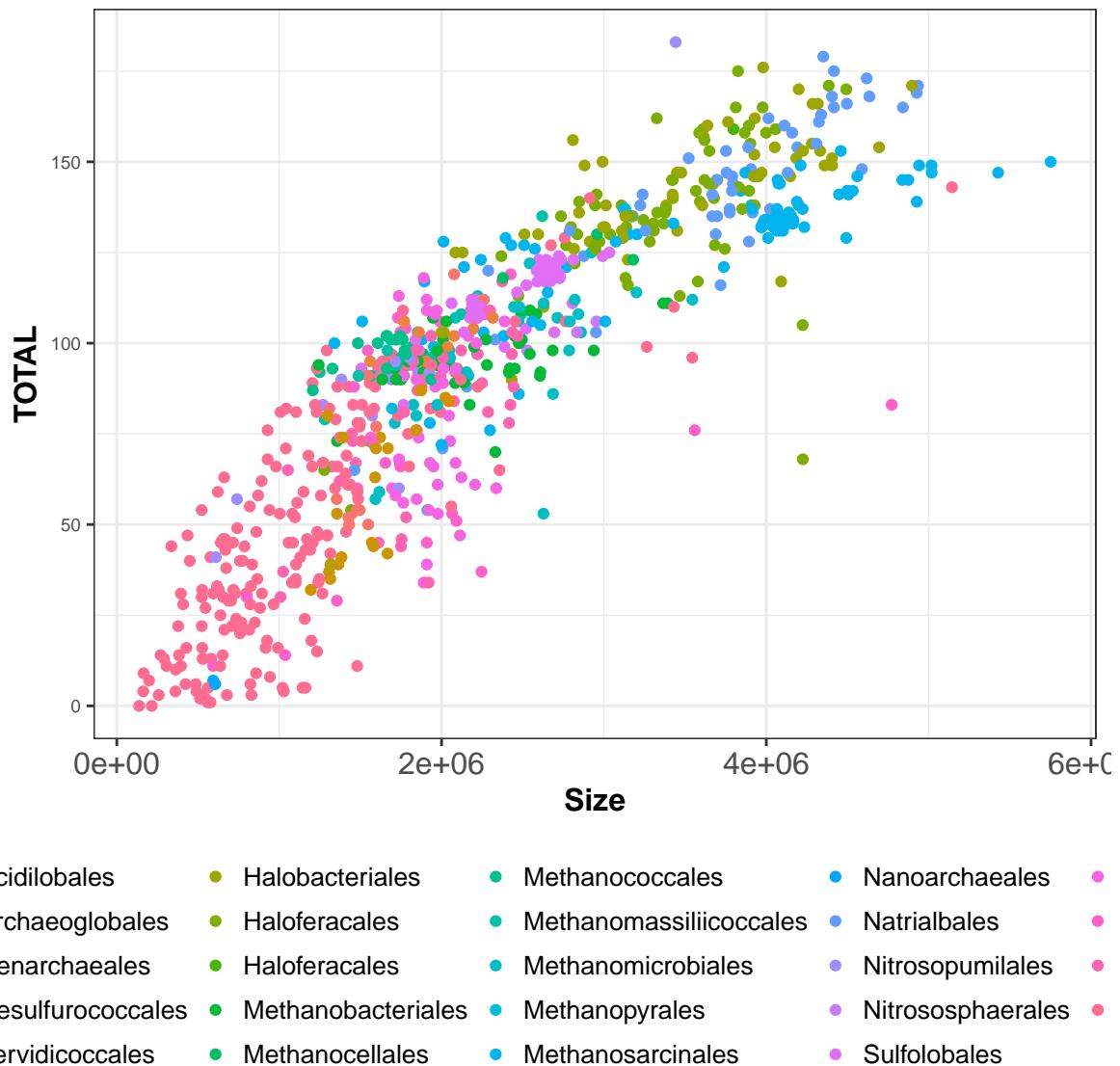


Figure 4.4: Correlation between Archaea genome size and central pathway expansions

Here is a reference to the size vs Total central pathway expansion plot: Figure 4.4.

Genome size vs Total central pathway expansion grided by order

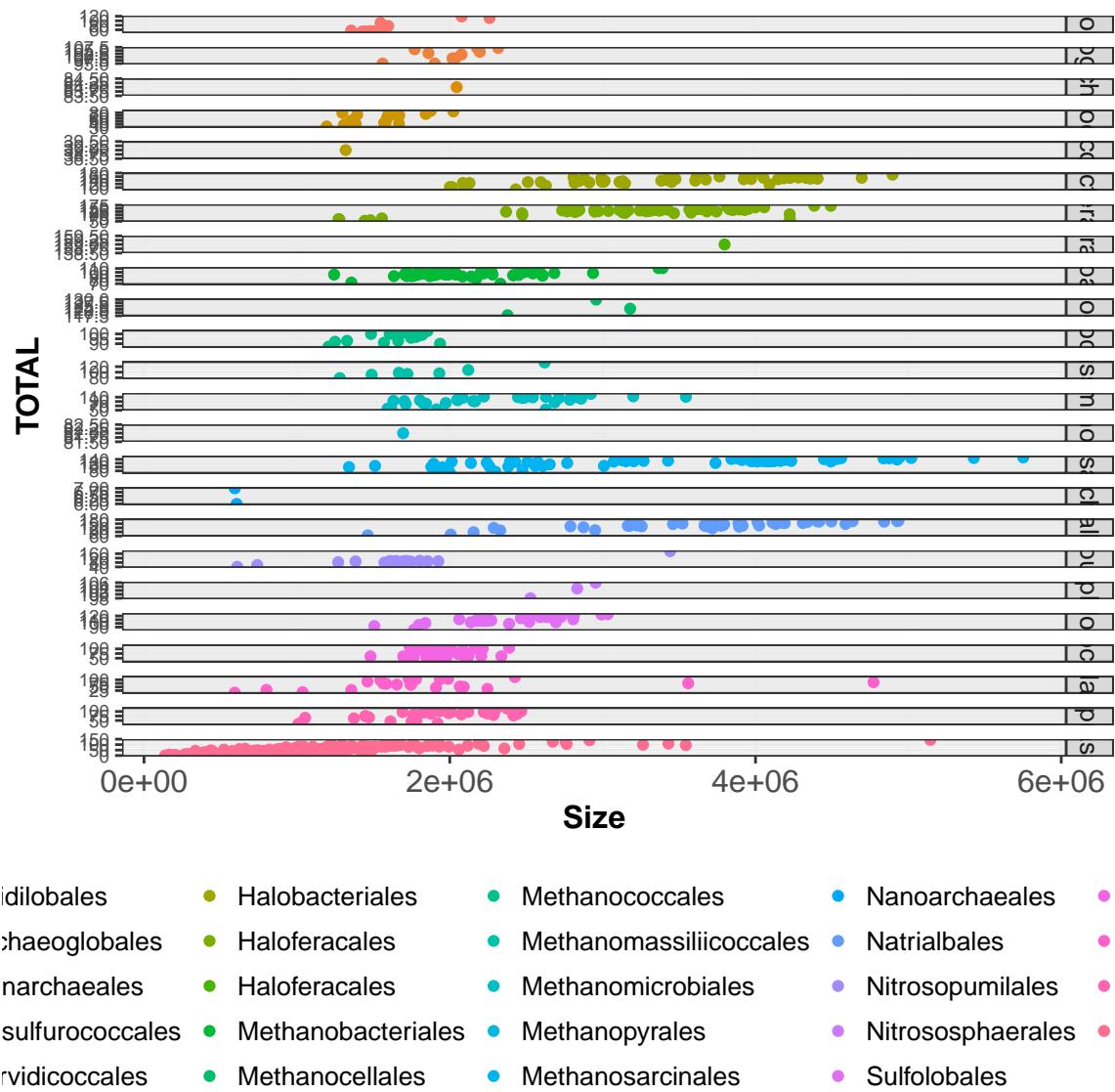


Figure 4.5: Correlation between Archaea genome size and central pathway expansions grided by order

Here is a reference to the Genome size vs Total central pathway expansion grided by order plot: Figure 4.5.

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

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

Warning: Removed 65604 rows containing missing values (geom\_point).

Genome size vs Total central pathway expansion coloured by metabolic Family

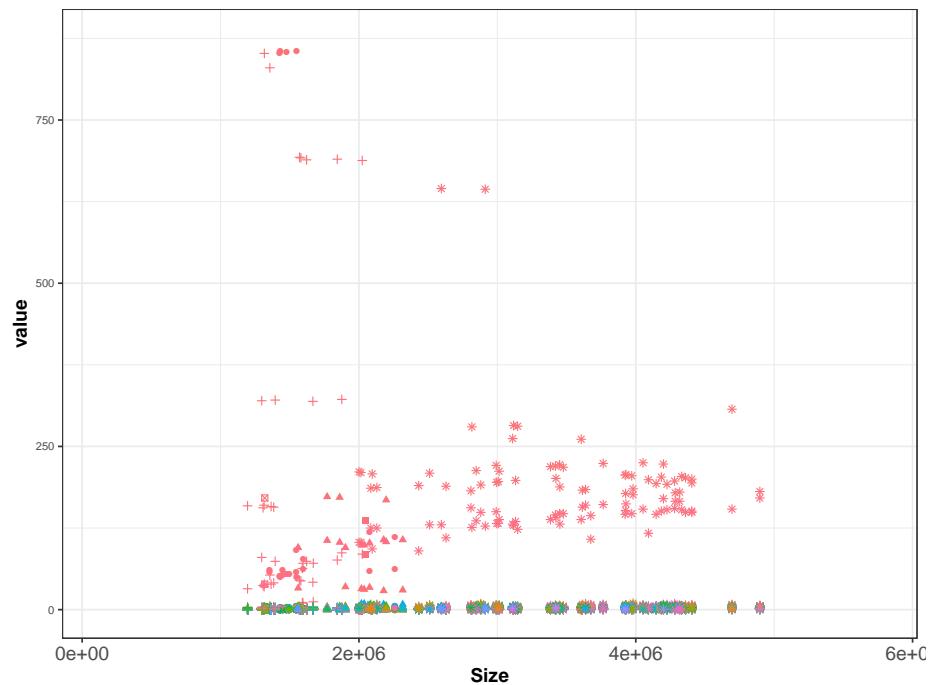


Figure 4.6: Correlation between Archaea Genome size vs Total central pathway expansion coloured by metabolic Family

Here is a reference to the Genome size vs Total central pathway expansion coloured by metabolic Family plot: Figure 4.6.

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: ??.

## 4.5 Natural products

#### 4.5.1 Natural products recruitments from EvoMining heat-plot

We can see natural products recruitment after central pathways expansions colored by their kingdom.

Natural products recruited by metabolic family, colored by phylogenetic origin.

## Recruitments after central pathways expansions coloured by Kingdom

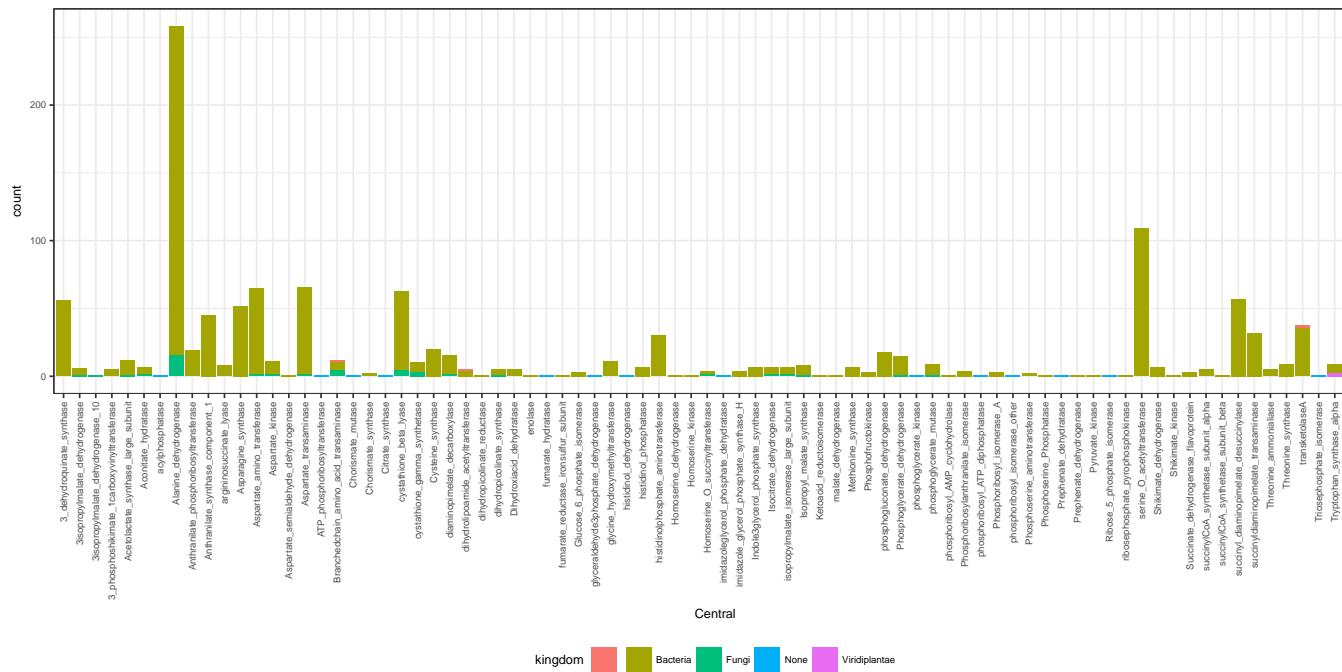


Figure 4.7: Archaeas Recruitmens on central families coloured by kingdom

Here is a reference to Recruitments after central pathways expansions coloured by Kingdom plot: Figure 4.7.

### Recruitments after central pathways expansions coloured by taxonomy

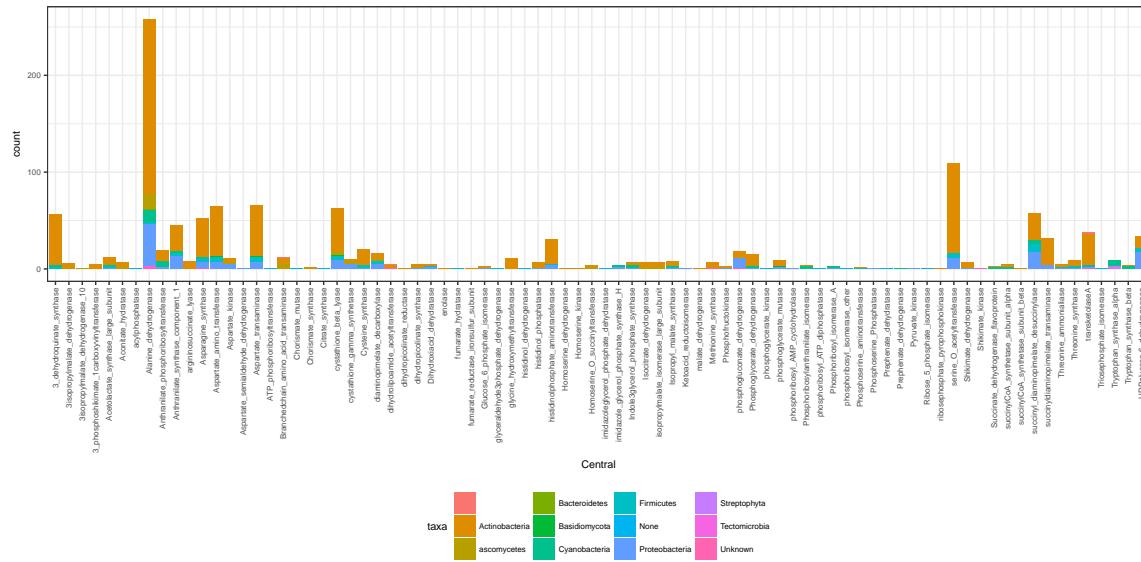


Figure 4.8: Archaeas Recruitmens on central families coloured by taxonomy

Here is a reference to Recruitments after central pathways expansions coloured by taxonomy plot: Figure 4.8.

## 4.6 Archaeas AntiSMASH

Taxonomical diversity on Archaeabacteria Data

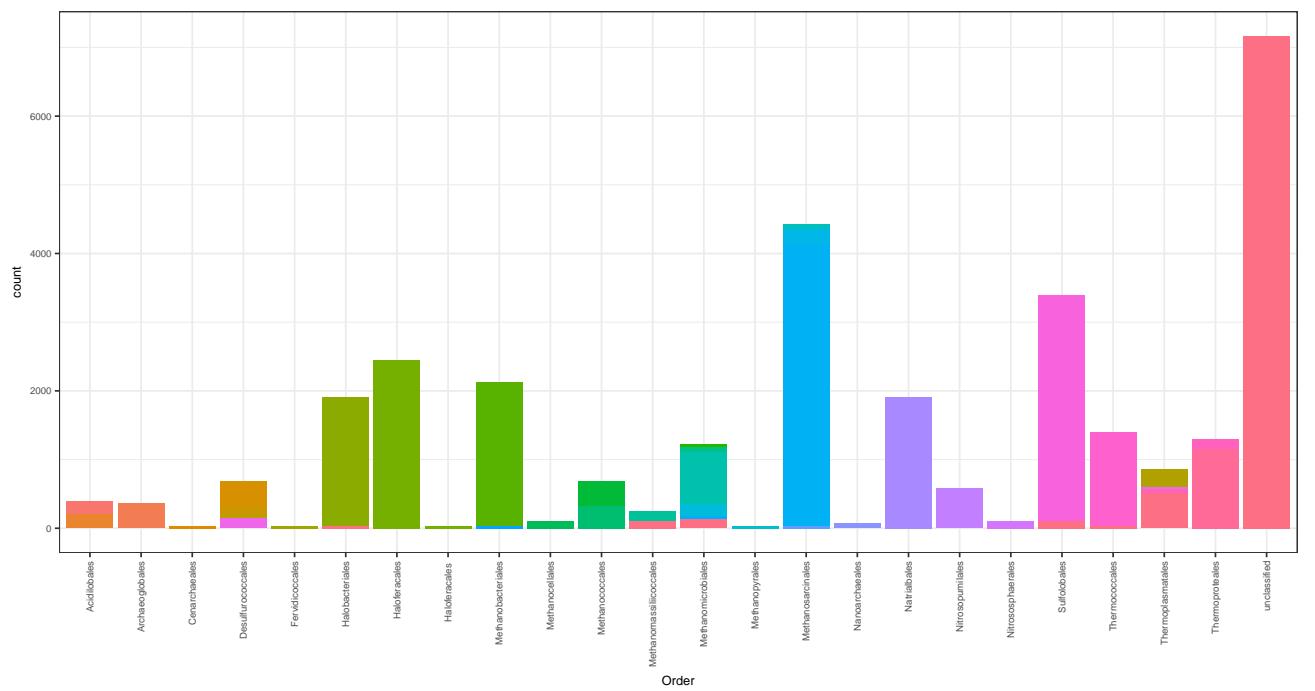


Figure 4.9: Archaeas Diversity

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

## Smash diversity

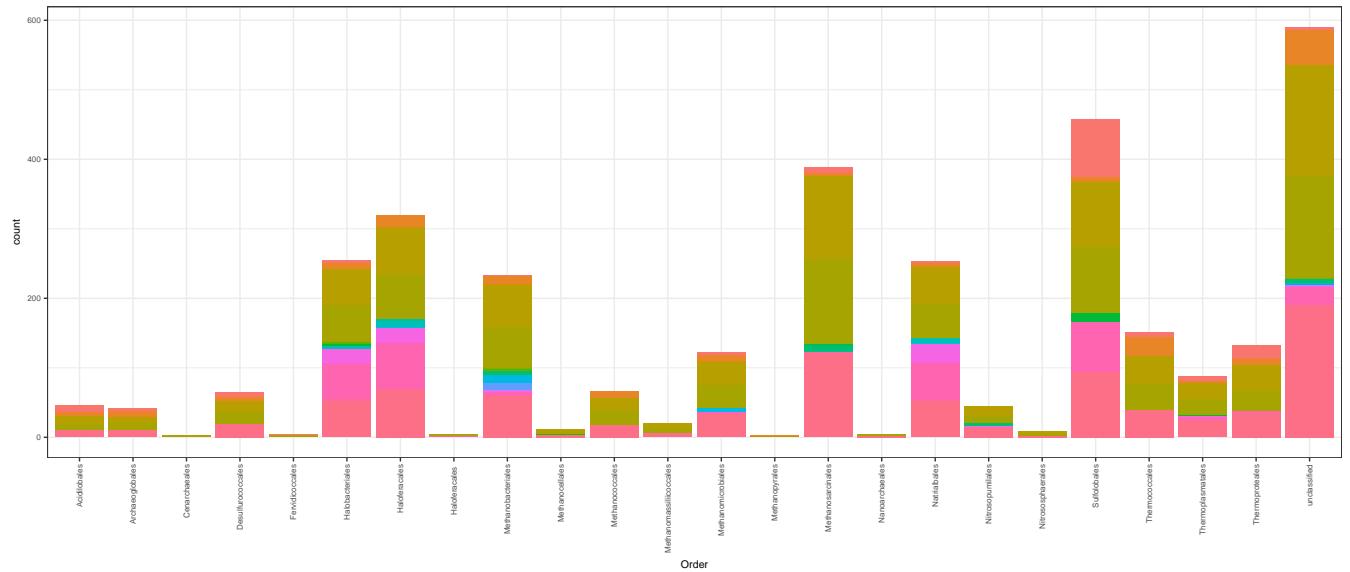


Figure 4.10: Archaeas Smash Taxonomical Diversity

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

### 4.6.1 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

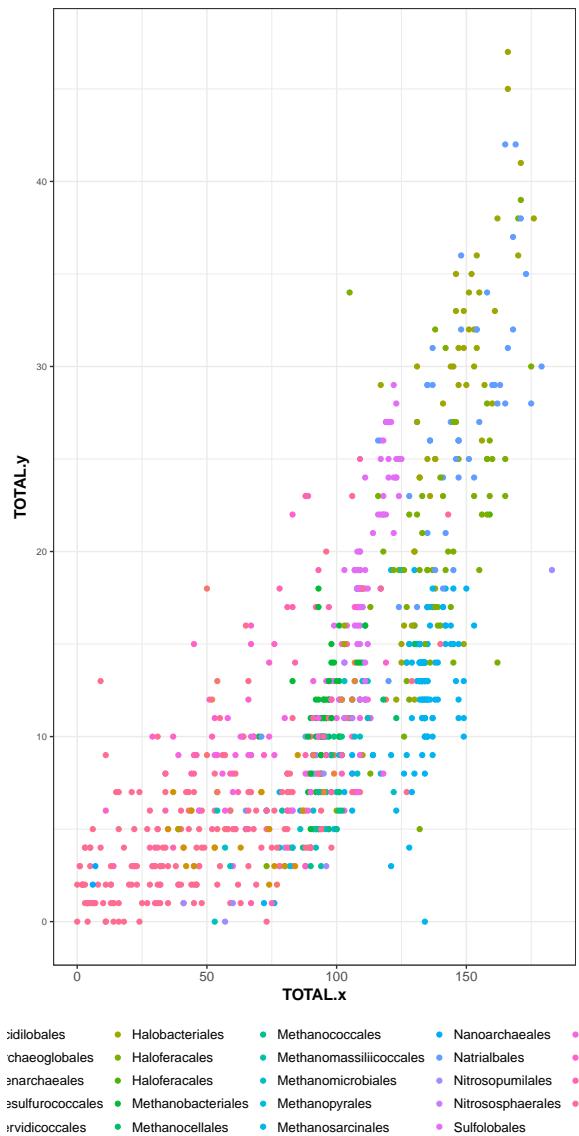


Figure 4.11: Correlation between Archaeas central pathway expansions and antismash Natural products detection

Here is a reference to the expansions vs antismash NP's clusters plot: Figure 4.11.

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

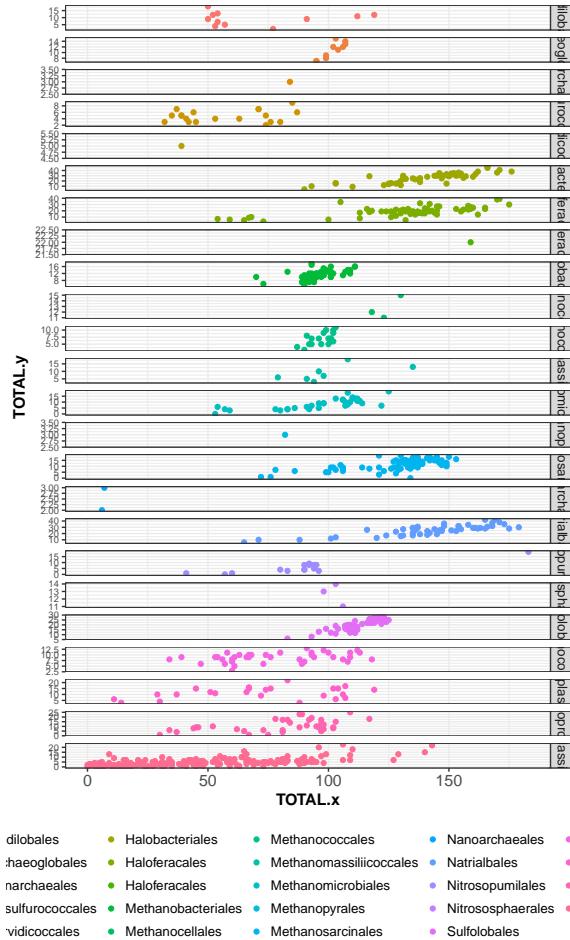


Figure 4.12: Correlation between Archaea's central pathway expansions and antismash NP's clusters splitted by order plot

Here is a reference to the expansions vs antismash NP's clusters splitted by order plot Figure 4.12.

## AntisMAsh vs Expansions by taxonomic Family

Natural products coloured by family

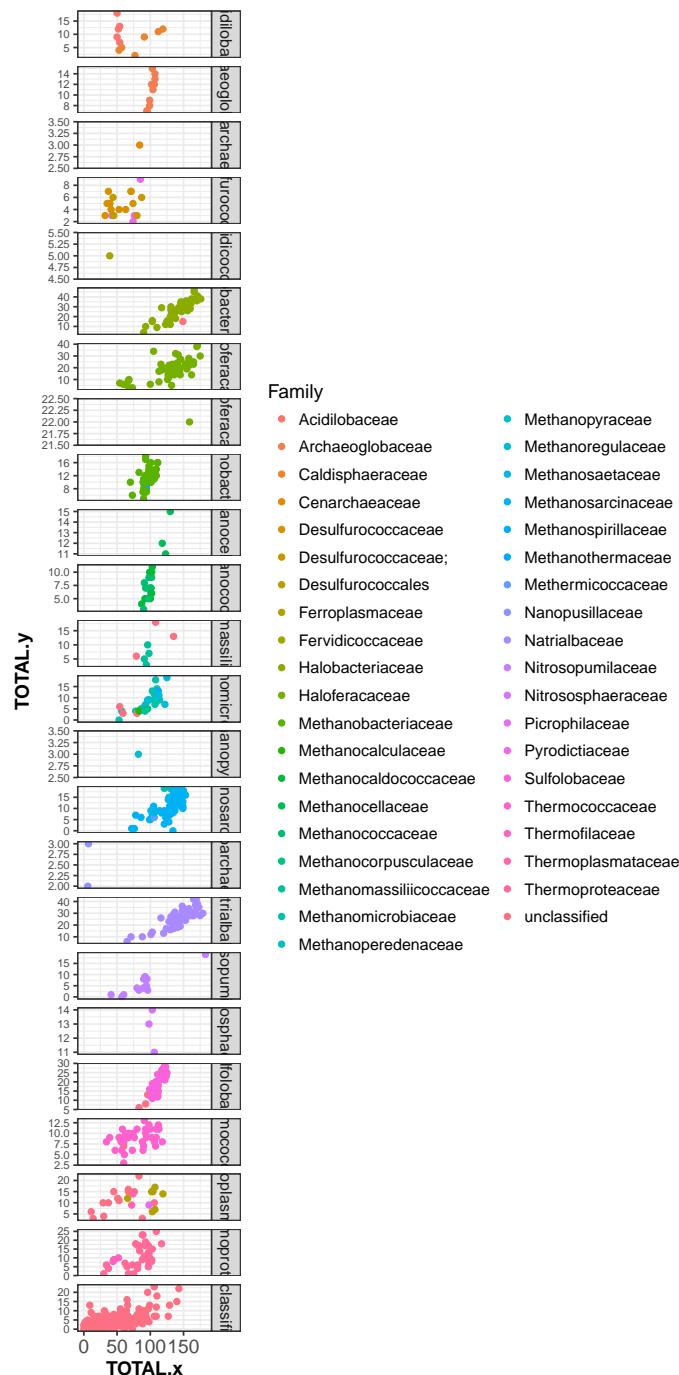


Figure 4.13: Archaeas Natural products by family

Here is a reference to the Natural products coloured by family plot Figure 4.13.

## 4.7 Selected trees from EvoMining

Phosphoribosyl\_isomerase\_3 family

Figure from EvoMining

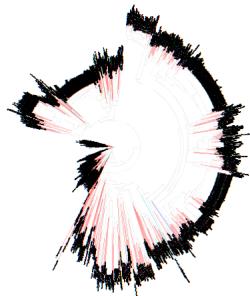


Figure 4.14: Phosphoribosyl isomerase A EvoMiningtree

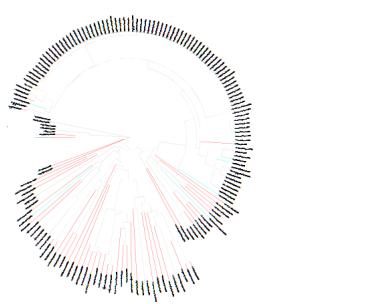


Figure 4.15: Phosphoribosyl isomerase other EvoMiningtree

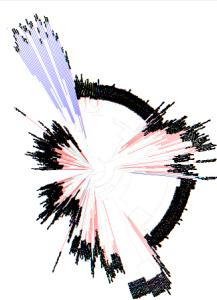


Figure 4.16: Phosphoribosyl anthranilate isomerase EvoMiningtree

## 4.8

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* [126] ## 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`.

## 4.9 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: [145]. 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>), *bibtextstyles* (about how to find and use the bibliography style that best suits your needs, at <http://web.reed.edu/cis/help/latex/bibtextstyles.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>[146]

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.”

## 4.10 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.

## 4.11 Actinobacteria

Actinobacteria is an ancient phylum {Referencia de luis}

## 4.12 Tables

Table 4.2: Correlation of Inheritance Factors for Parents and Child

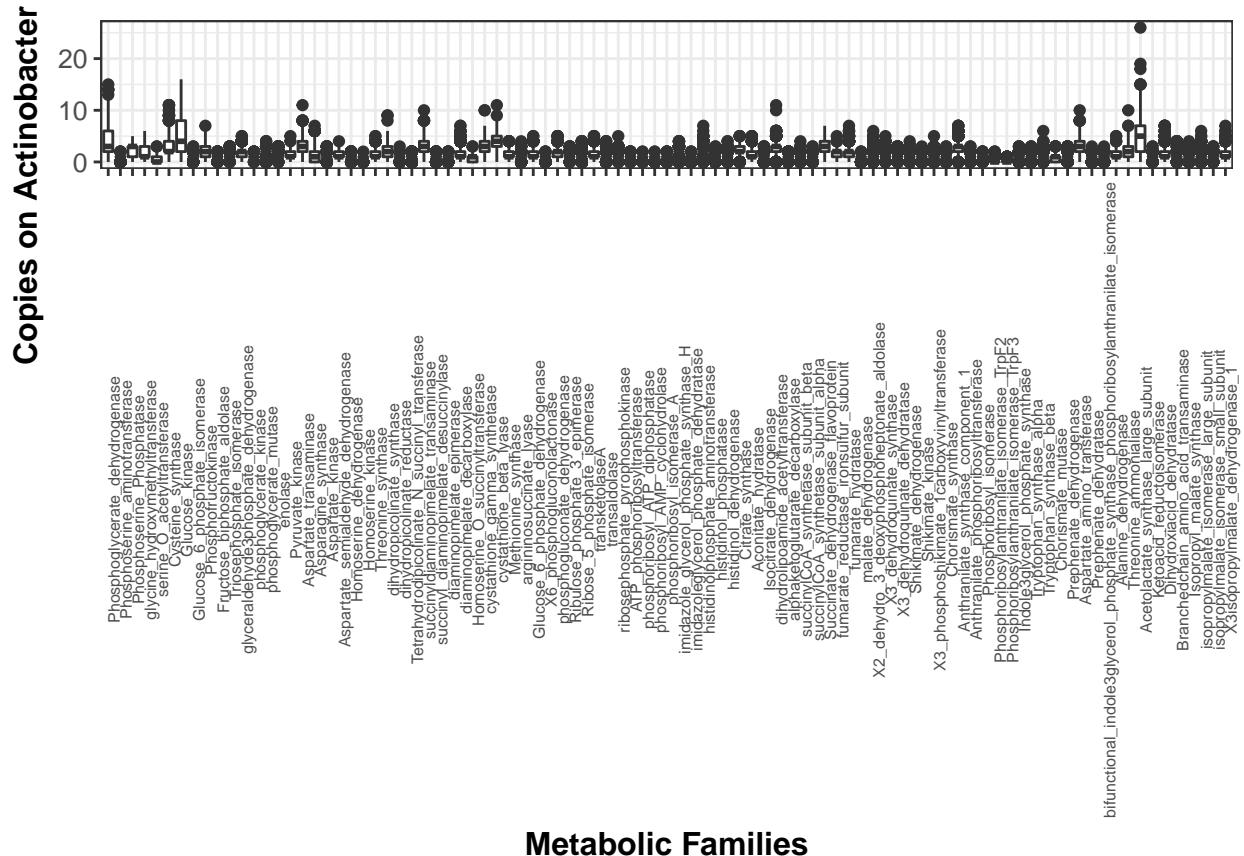
Factors	Correlation between Parents & Child
GenomeDB	1245
Families	65

<sup>3</sup>[147]



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

#### 4.12.1 Expansions BoxPlot by metabolic family



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

Here is a reference to the expansion boxplot: Figure 4.17.

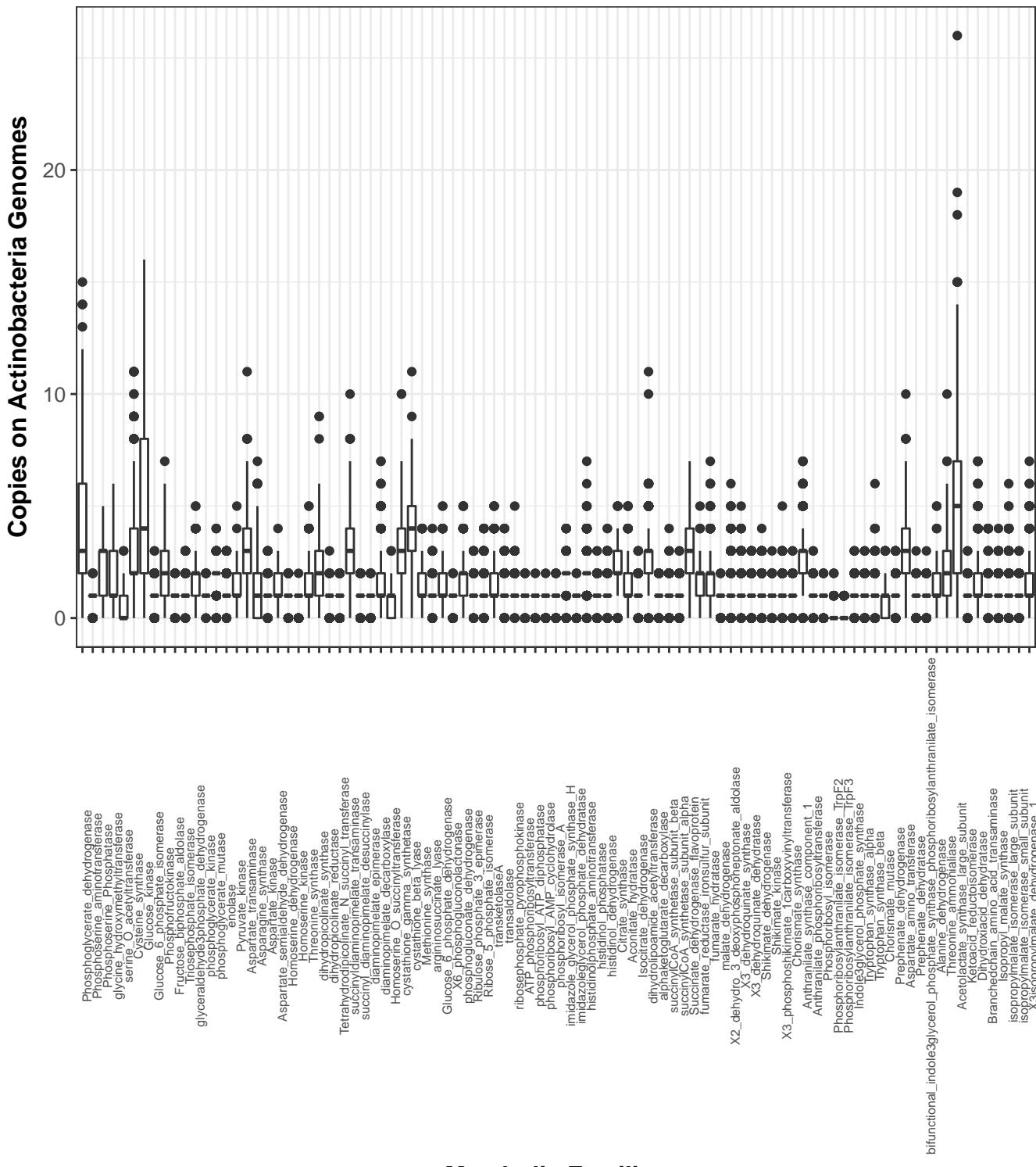


Figure 4.17: Expansions Boxplot

## 4.13 Central pathway expansions

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

Here is a reference to the HeatPlot: Figure 4.18.



PPP pahtway expansions restricted to *Streptomycetaceae* family HeatPlot: Figure 4.18.

Here is a reference to the HeatPlot: Figure 4.19.

288310	
288306	
288302	
288298	
288289	
288280	
288278	
288277	
288268	
288267	
288266	
288261	
288254	
288233	
288231	
288218	
288175	
288188	
288178	
288152	
288032	
288019	
288012	
287996	
287981	
287979	
287965	
287955	
287953	
287950	
287948	
262855	
252178	
252176	
252172	
252171	
252170	
252168	
252167	
252165	
242654	
242652	
242564	
242541	
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	
242462	
242461	
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	
242343	
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	

## 4.14 Genome Size correlations

### 4.14.1 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

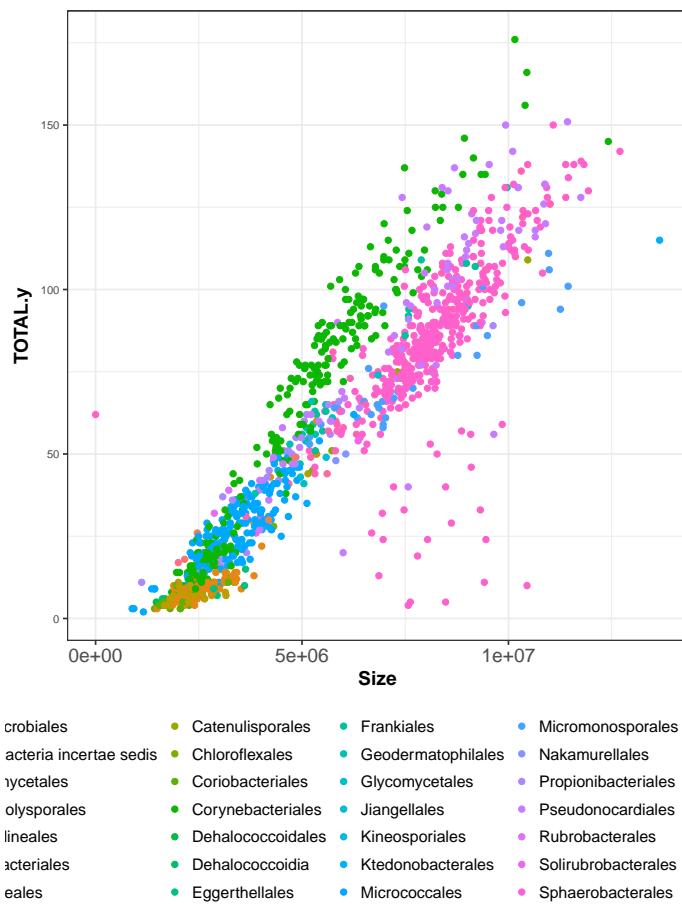


Figure 4.20: Correlation between Actinos genome size and antismash Natural products detection colored by Order

Here is a reference to Genome size vs Total antismash cluster: Figure 4.20.

Genome size vs Total antismash cluster detected splitted by order

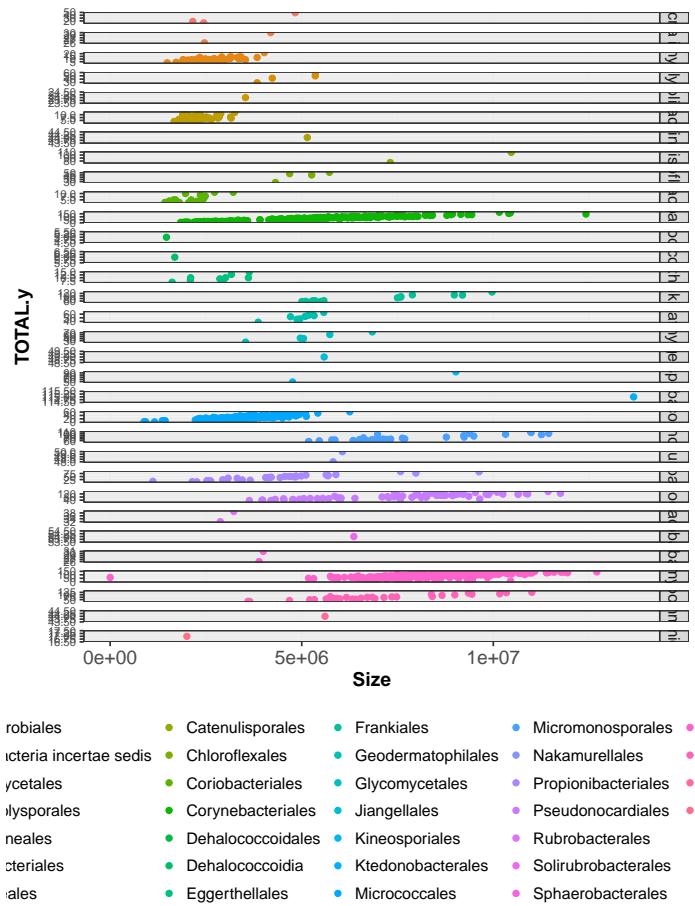


Figure 4.21: Correlation between Actinos genome size and antismash Natural products detection grided by Order

Here is a reference to Correlation between genome size and antismash Natural products detection grided by Order plot: Figure 4.21.

#### 4.14.2 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

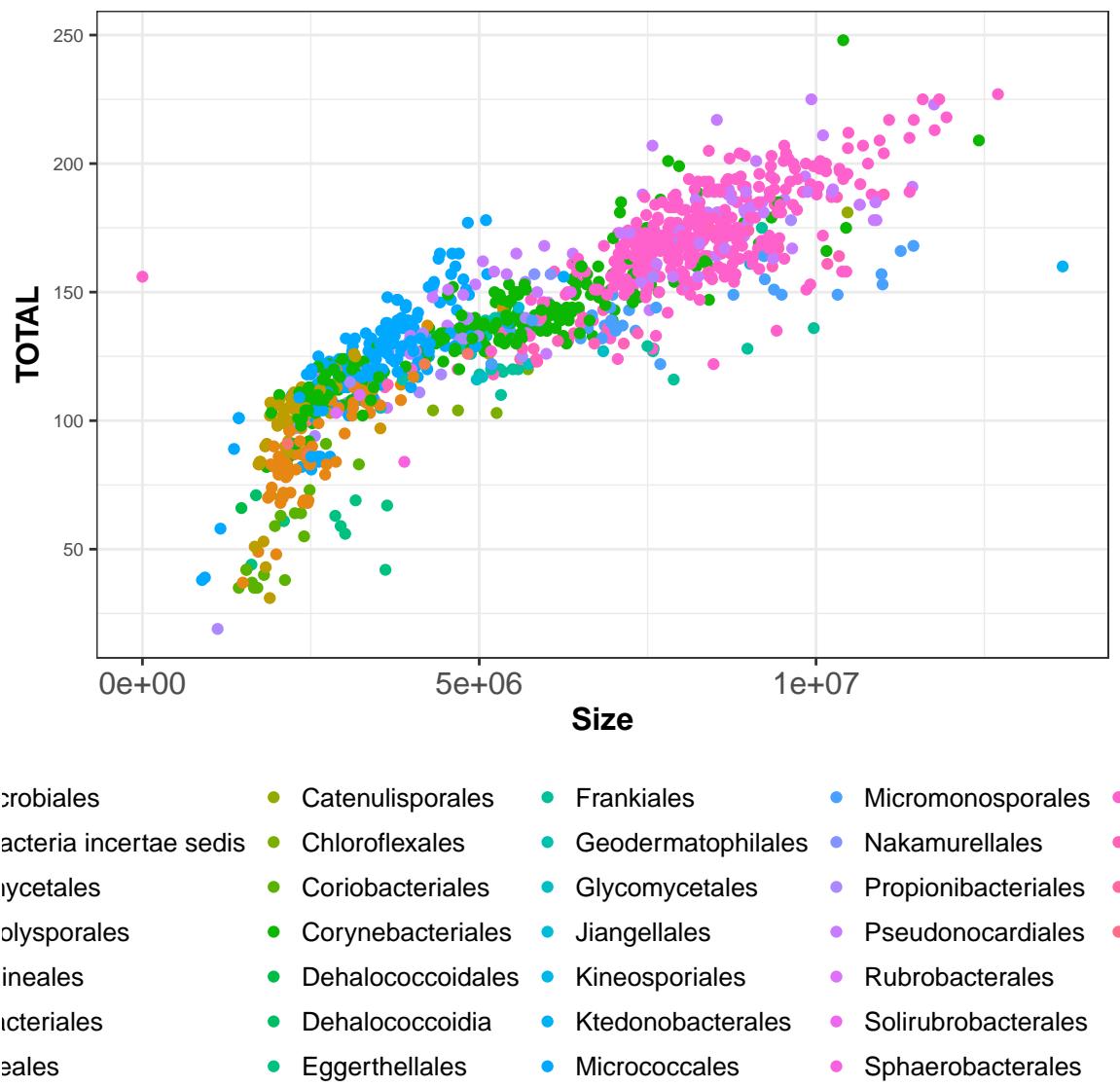


Figure 4.22: Correlation between Actinos genome size and central pathway expansions

Here is a reference to the size vs Total central pathway expansion plot: Figure 4.22.

Genome size vs Total central pathway expansion grided by order

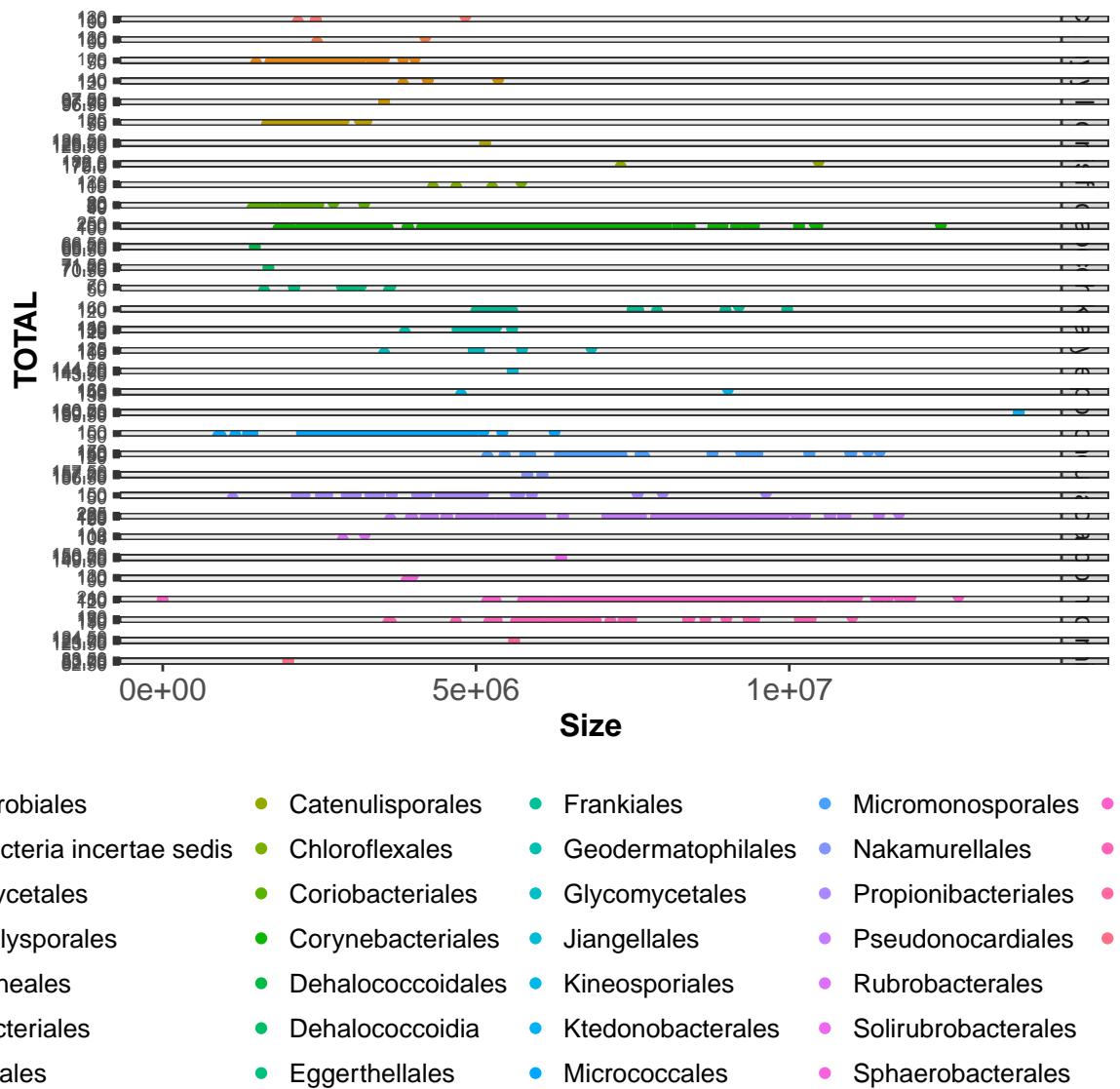


Figure 4.23: Correlation between Actinos genome size and central pathway expansions grided by order

Here is a reference to the Genome size vs Total central pathway expansion grided by order plot: Figure 4.23.

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

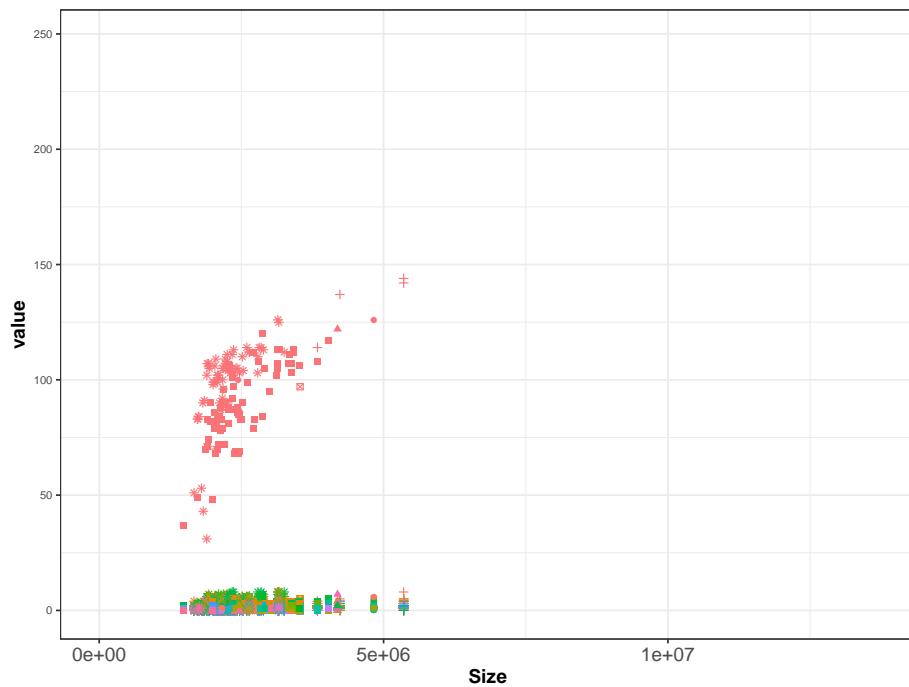


Figure 4.24: Correlation between Actinos Genome size vs Total central pathway expansion coloured by metabolic Family

Here is a reference to the Genome size vs Total central pathway expansion coloured by metabolic Family plot: Figure 4.24.

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: ??.

## 4.15 Natural products

#### 4.15.1 Natural products recruitments from EvoMining heat-plot

We can see natural products recruitment after central pathways expansions colored by their kingdom.

Natural products recruited by metabolic family, colored by phylogenetic origin.

## Recruitments after central pathways expansions coloured by Kingdom

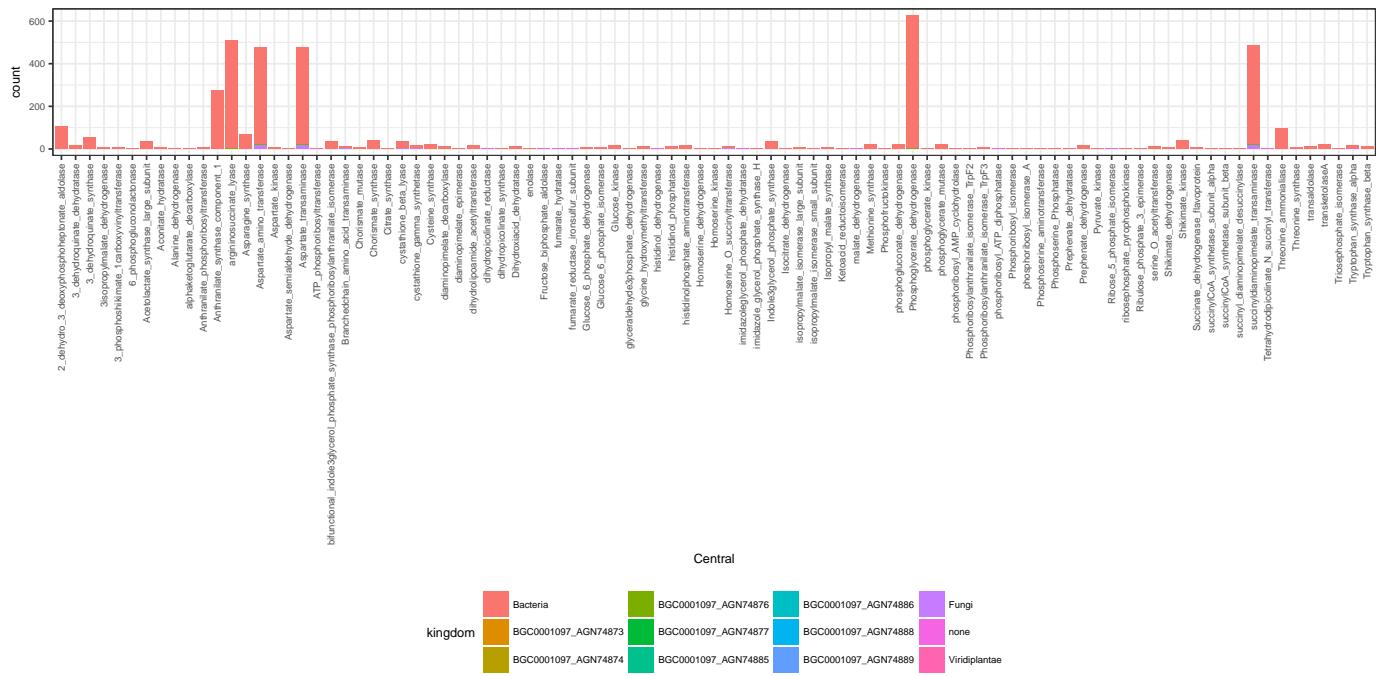


Figure 4.25: Actinos Recruitmens on central families coloured by kingdom

Here is a reference to Recruitments after central pathways expansions coloured by Kingdom plot: Figure 4.25.

## Recruitments after central pathways expansions coloured by taxonomy

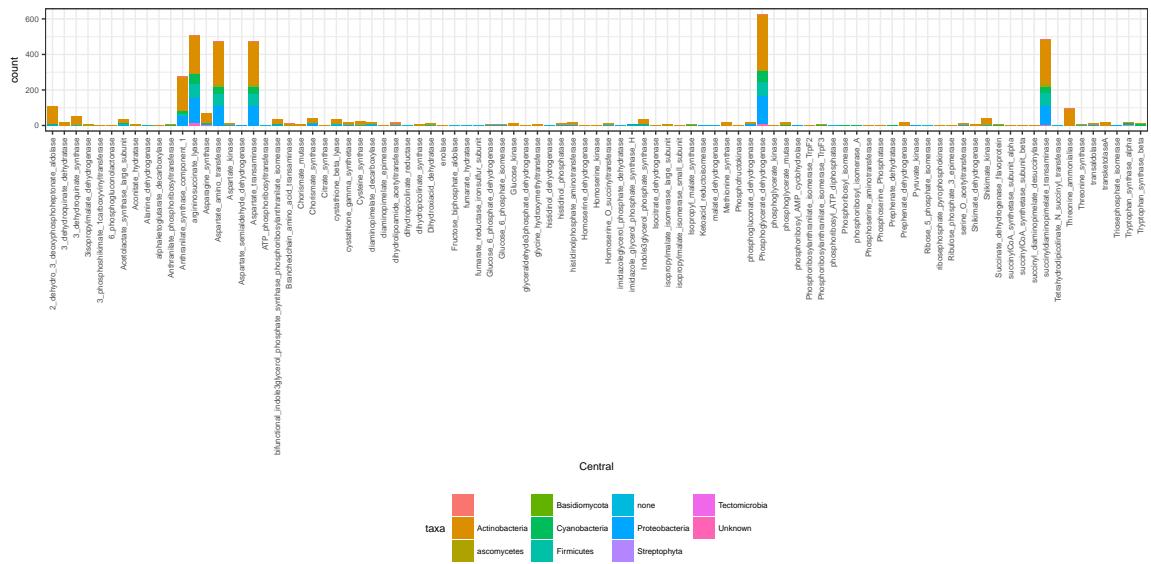


Figure 4.26: Actinos Recruitmens on central families coloured by taxonomy

Here is a reference to Recruitments after central pathways expansions colourd by taxa plot: Figure 4.26.

## 4.16 Actinos AntiSMASH

Taxonomical diversity on Actinosbacteria Data

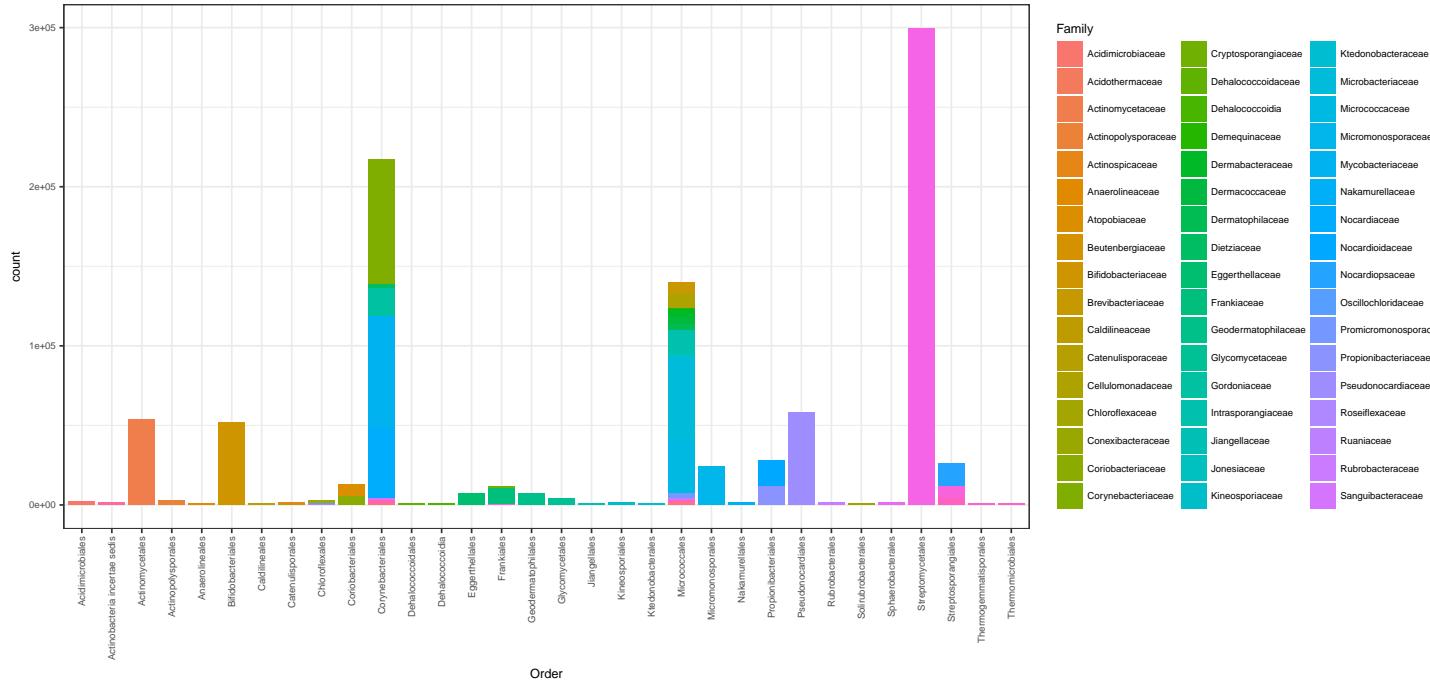


Figure 4.27: Actinos Diversity

Here is a reference to Recruitments after central pathways expansions coloured by taxa plot: Figure 4.27.

## Smash diversity

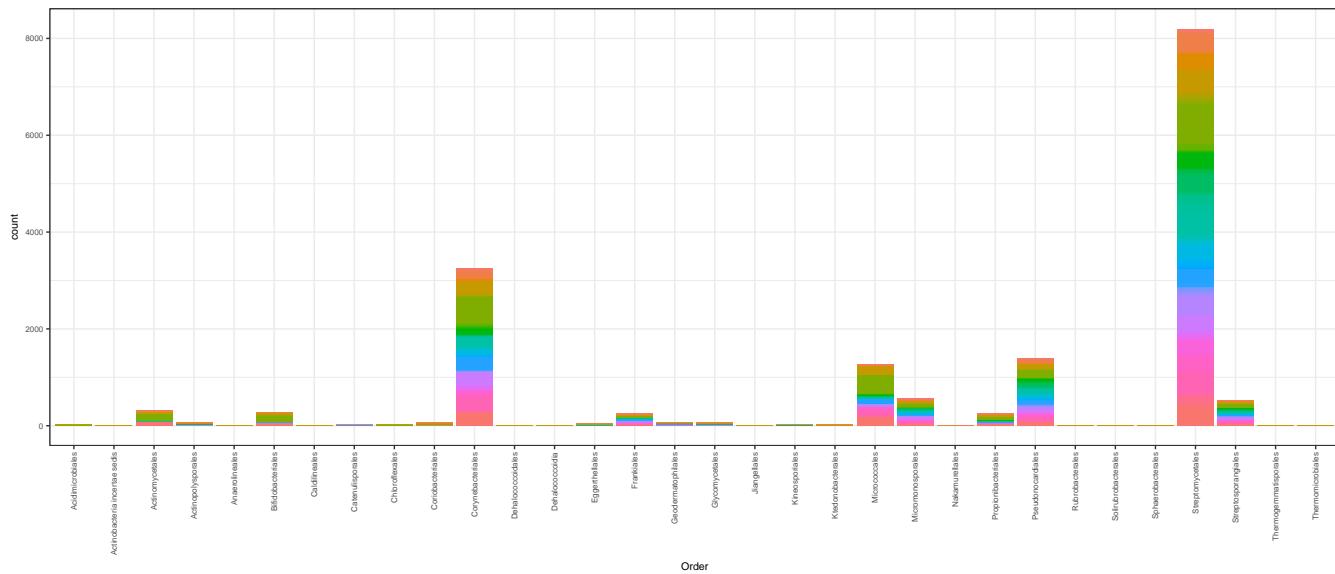


Figure 4.28: Actinos Smash Taxonomical Diversity

Here is a reference to Recruitments after central pathways expansions colourd by taxa plot: Figure 4.28.

### 4.16.1 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 4.29: Correlation between Actinos central pathway expansions and antismash Natural products detection

Here is a reference to the expansions vs antismash NP's clusters plot: Figure 4.29.

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

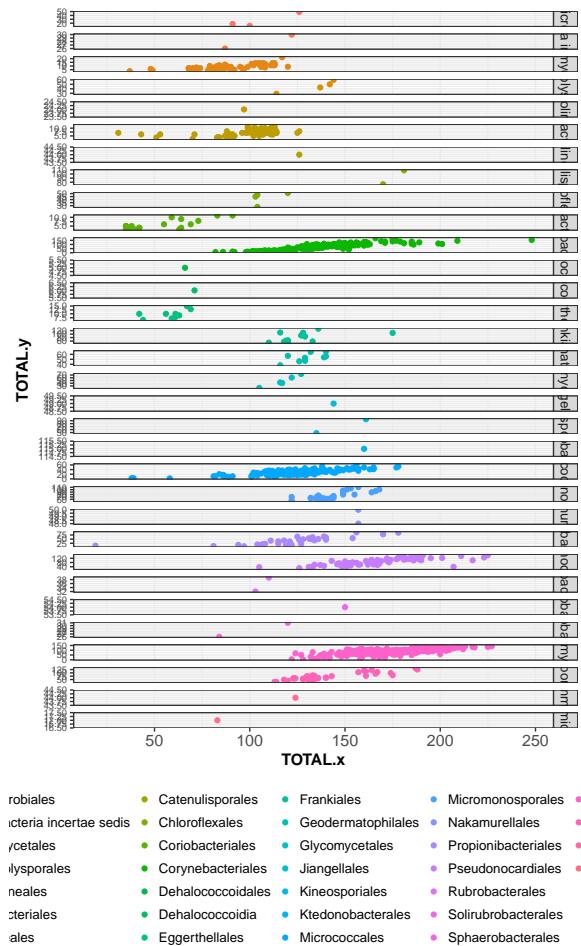


Figure 4.30: Correlation between Actinos central pathway expansions and antismash Natural products detection

Here is a reference to the expansions vs antismash NP's clusters splitted by order plot Figure 4.30.

AntisMAsh vs Expansions by taxonomic Family

Natural products colured by family



Figure 4.31: Actinos Natural products by family

Here is a reference to the Natural products colured by family plot Figure 4.31.

## 4.17 Selected trees from EvoMining

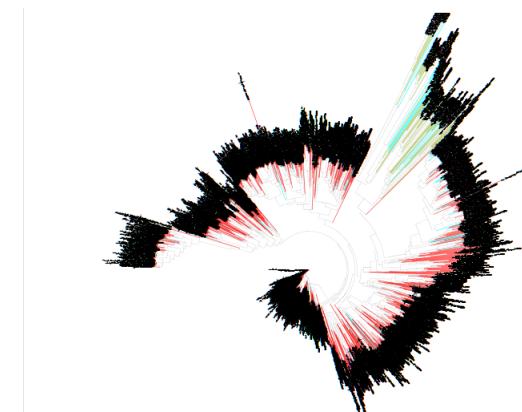


Figure 4.32: Enolase EvoMiningtree

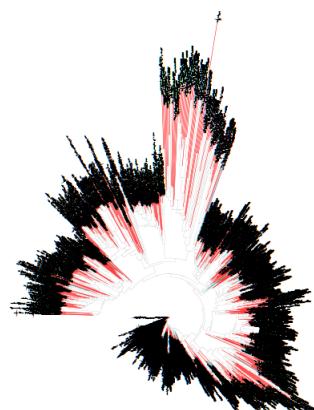


Figure 4.33: Phosphoribosyl isomerase EvoMiningtree

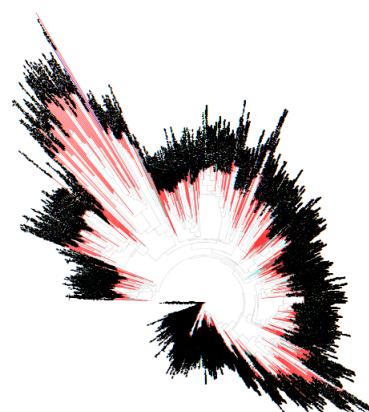


Figure 4.34: Phosphoribosyl isomerase A EvoMiningtree

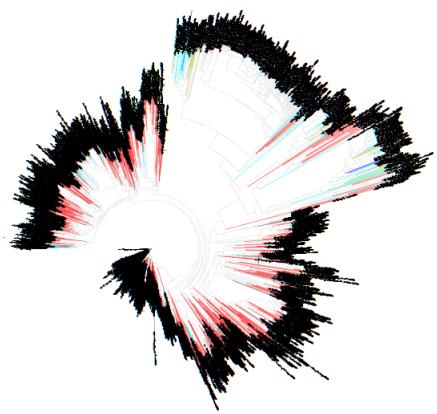


Figure 4.35: phosphoshikimate carboxyvinyltransferase EvoMiningtree

## 4.18 Cyanobacteria

Cyanobacteria phylum {Referencia}

Cyanobacteria is a photosynthetic phylum that inhabits a broad range of habitats. The broad adaptive potential is on part driven by gene-family enlargement [133] 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 [134] Example of a PriA cluster toxins[103]

Fossil record situates Cyanobacteria [134] Molecular record and metabolic properties at [137]

## 4.19 Tables

Table 4.3: Families on Cyanobacteria

Factors	Correlation between Parents & Child
GenomeDB	1245
Families	65

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

#### 4.19.1 Expansions BoxPlot by metabolic family

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

Here is a reference to the expansion boxplot: Figure 4.36.

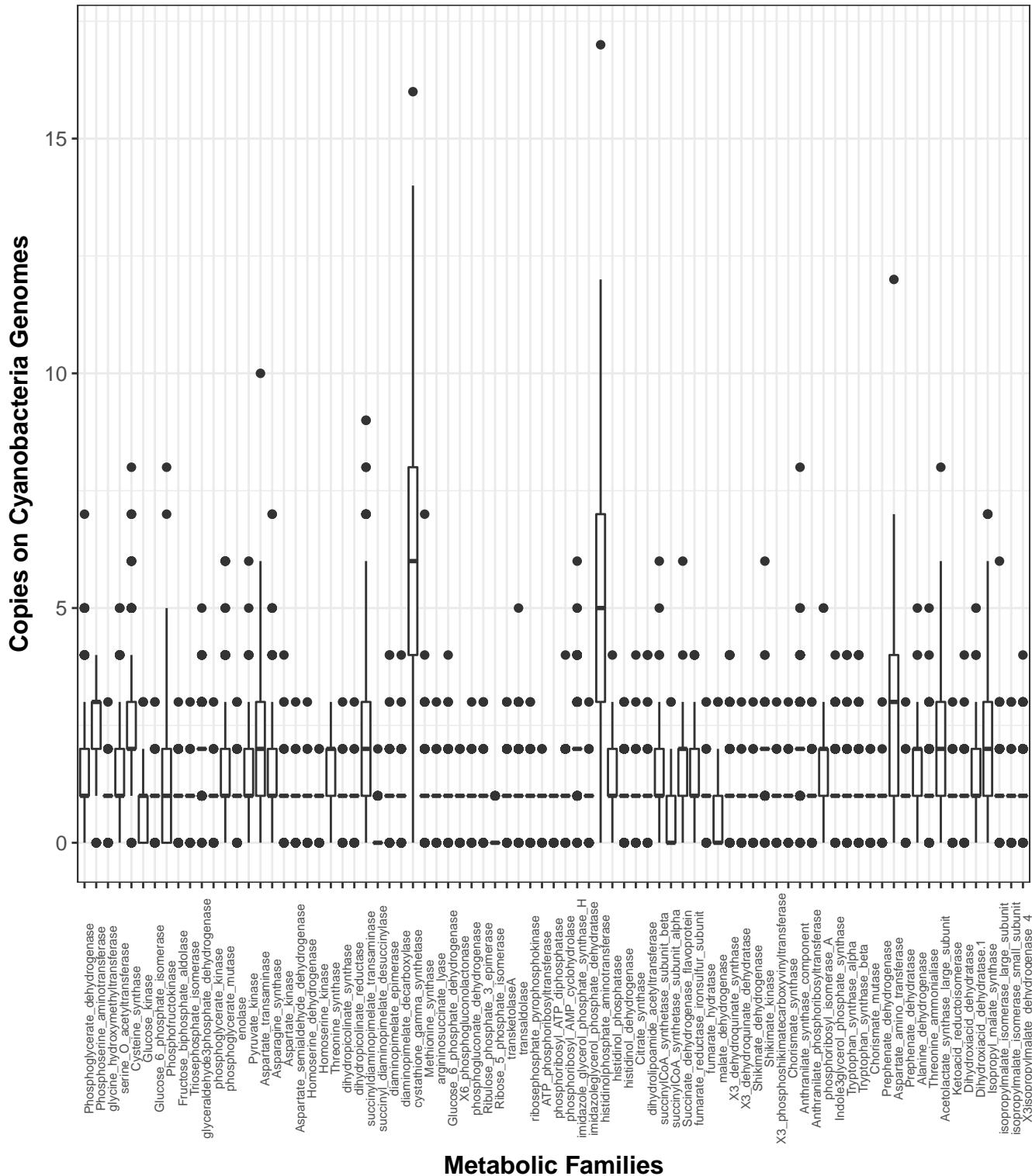


Figure 4.36: Expansions Boxplot

## 4.20 Central pathway expansions

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

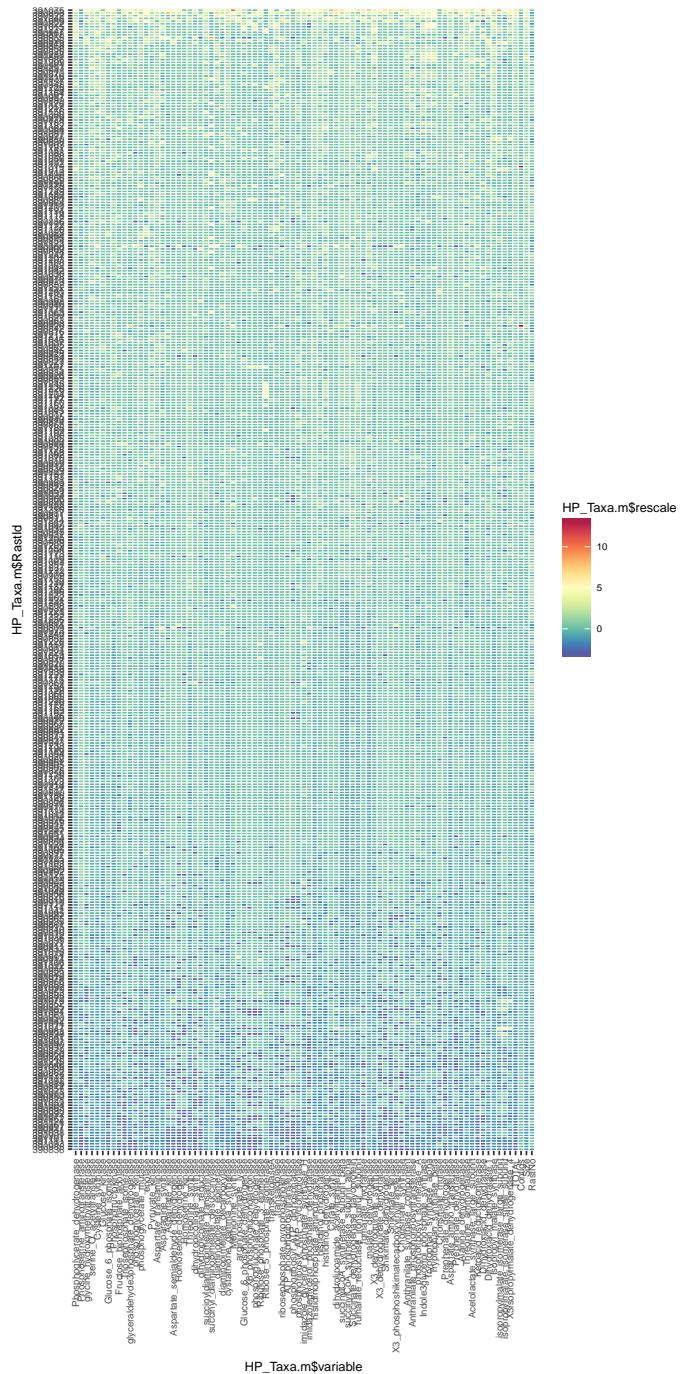


Figure 4.37: Cyanobacterial Heatplot

Here is a reference to the HeatPlot: Figure 4.37.

## 4.21 Genome Size correlations

### 4.21.1 Correlation between genome size and AntiSMASH products

Genome size vs Total antismash cluster coloured by order

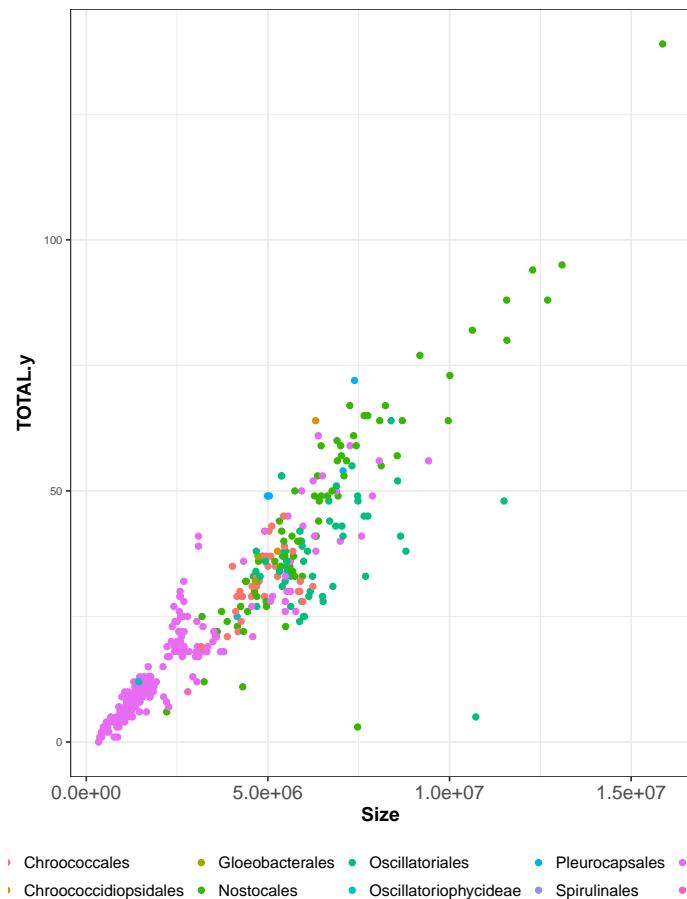


Figure 4.38: Correlation between genome size and antismash Natural products detection colored by Order

Here is a reference to Genome size vs Total antismash cluster: Figure 4.38.

Genome size vs Total antismash cluster detected splitted by order

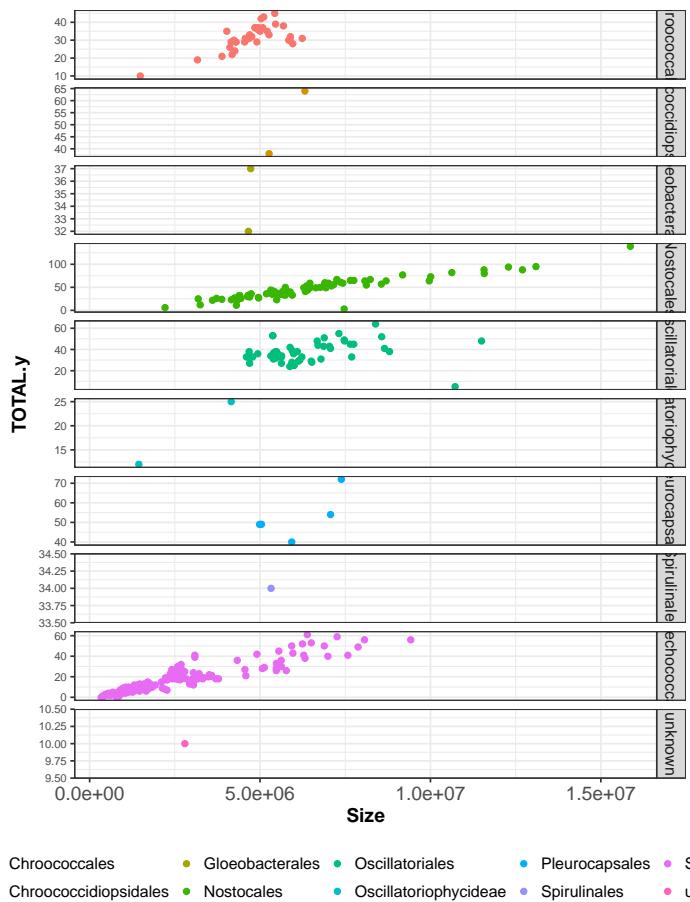


Figure 4.39: Correlation between genome size and antismash Natural products detection grided by Order

Here is a reference to Correlation between genome size and antismash Natural products detection grided by Order plot: Figure 4.39.

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

Genome size vs Total central pathway expansion coloured by order

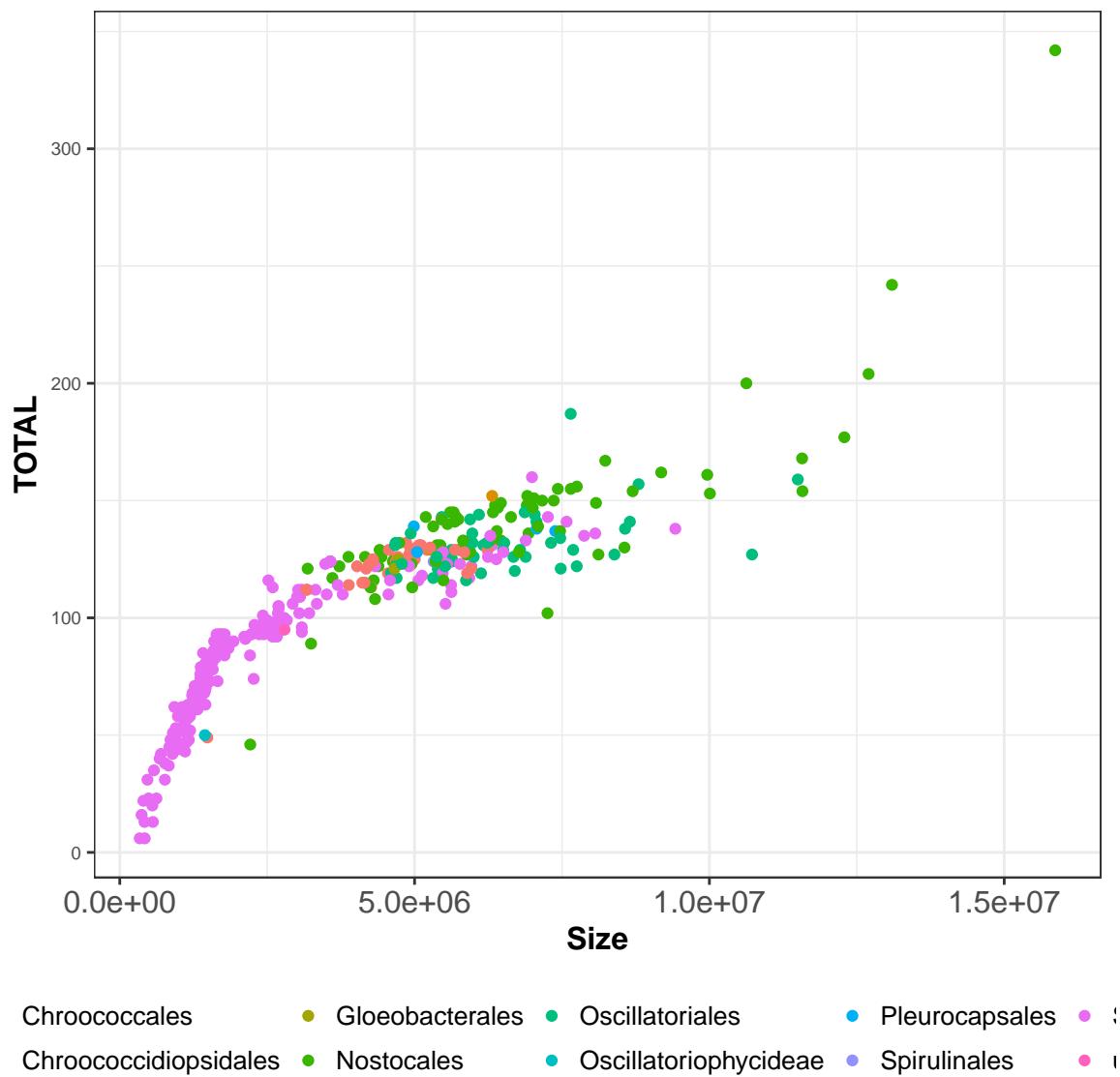


Figure 4.40: Correlation between genome size and central pathway expansions

Here is a reference to the size vs Total central pathway expansion plot: Figure 4.40.

Genome size vs Total central pathway expansion grided by order

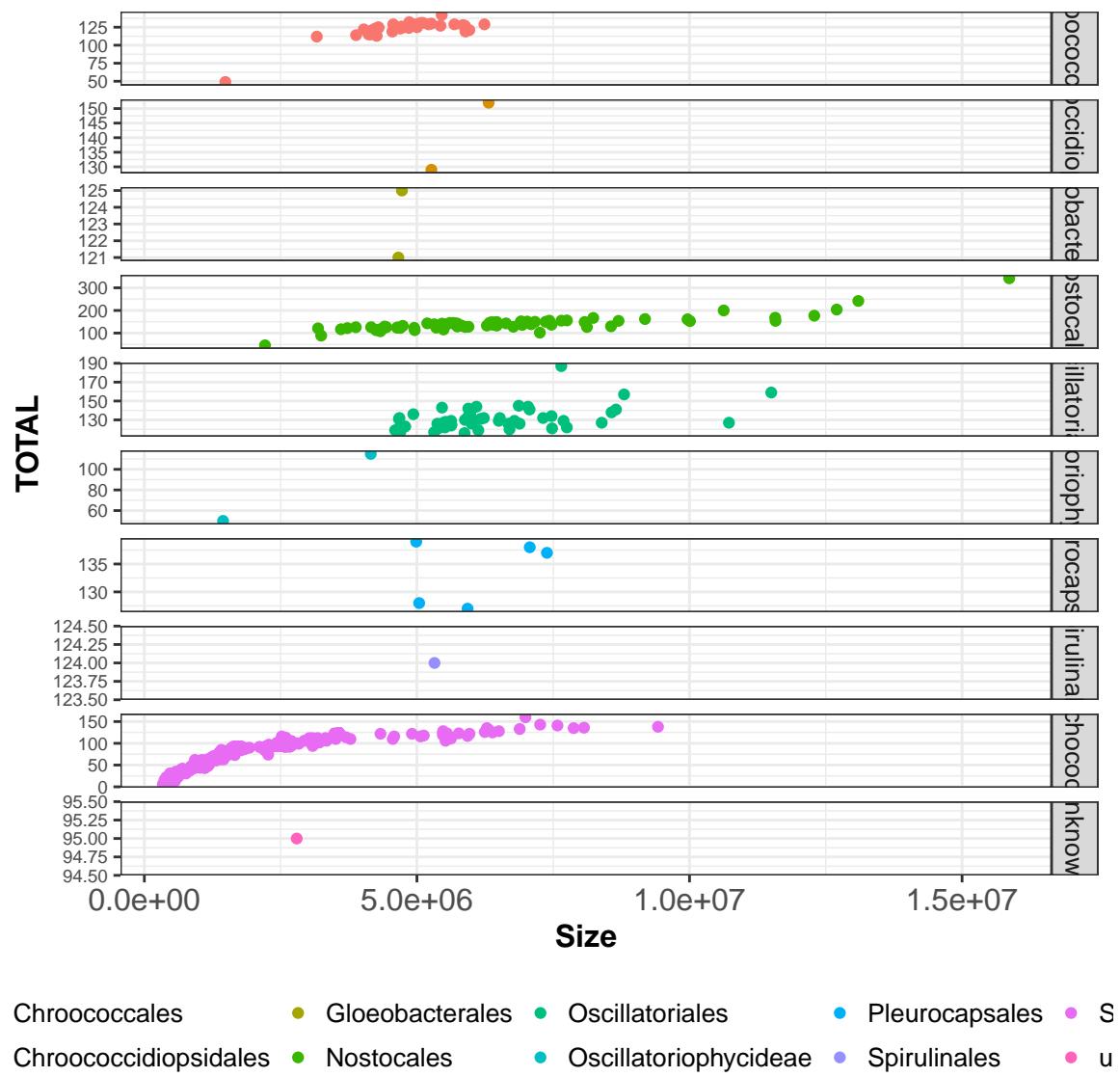


Figure 4.41: Correlation between genome size and central pathway expansions grided by order

Here is a reference to the Genome size vs Total central pathway expansion grided by order plot: Figure 4.41.

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

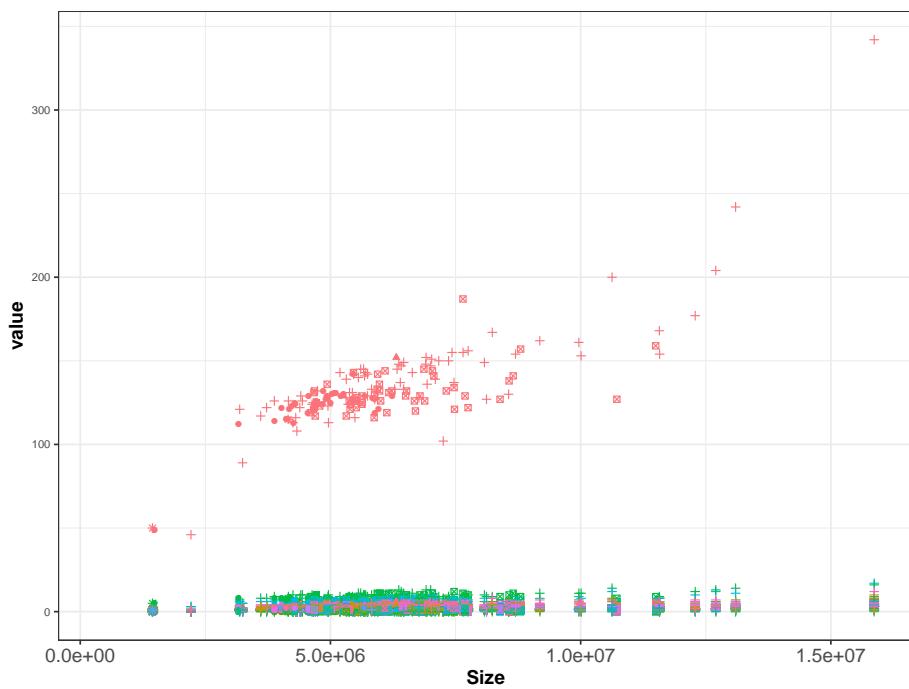


Figure 4.42: Correlation between Genome size vs Total central pathway expansion coloured by metabolic Family

Here is a reference to the Genome size vs Total central pathway expansion coloured by metabolic Family plot: Figure 4.42.

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: ??.

## 4.22 Natural products

#### 4.22.1 Natural products recruitments from EvoMining heat-plot

We can see natural products recruitment after central pathways expansions colored by their kingdom.

Natural products recruited by metabolic family, colored by phylogenetic origin.

## Recruitments after central pathways expansions coloured by Kingdom

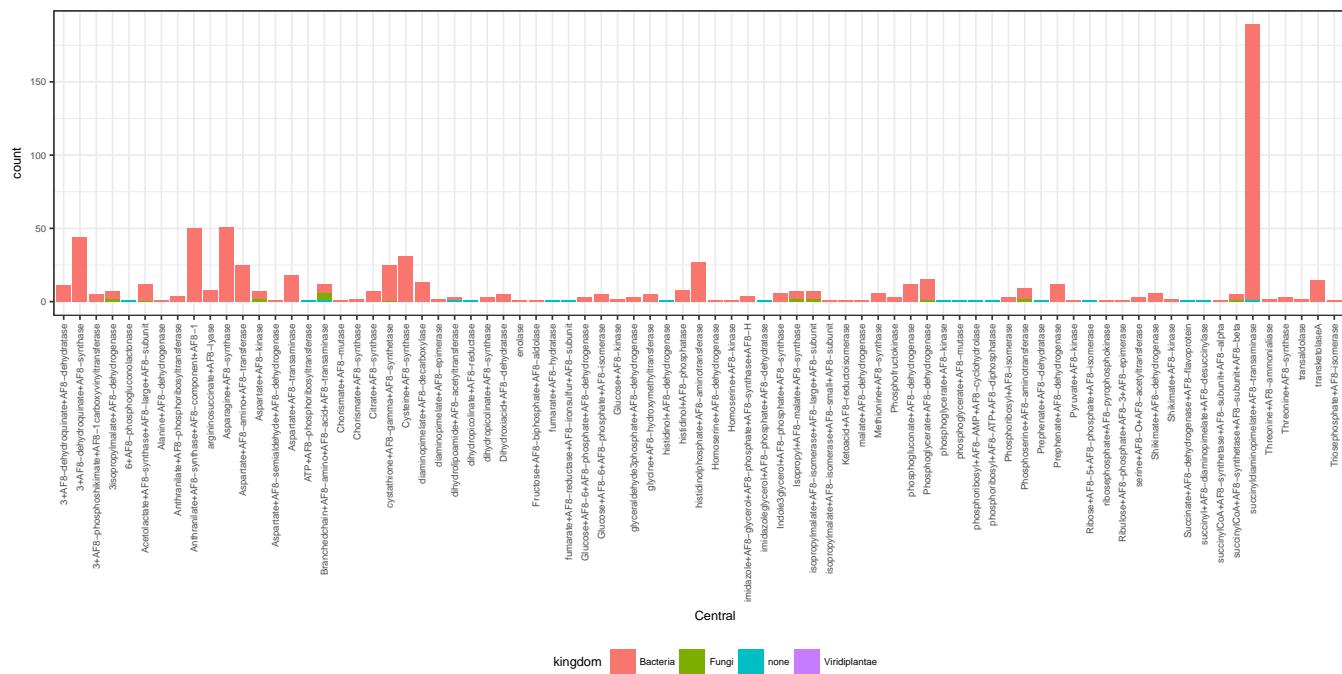


Figure 4.43: Recruitmens on central families coloured by kingdom

Here is a reference to Recruitments after central pathways expansions coloured by Kingdom plot: Figure 4.43.

## Recruitments after central pathways expansions coloured by taxonomy

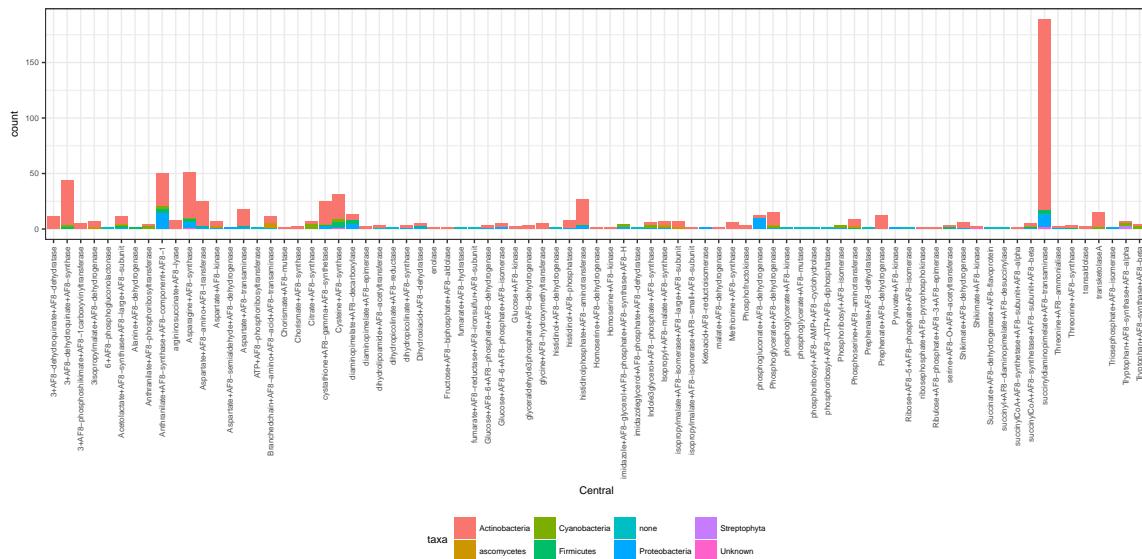


Figure 4.44: Recruitmens on central families coloured by taxonomy

Here is a reference to Recruitments after central pathways expansions colourd by taxa plot: Figure 4.44.

## 4.23 Cyanobacterias AntiSMASH

Taxonomical diversity on Cyanobacteria Data

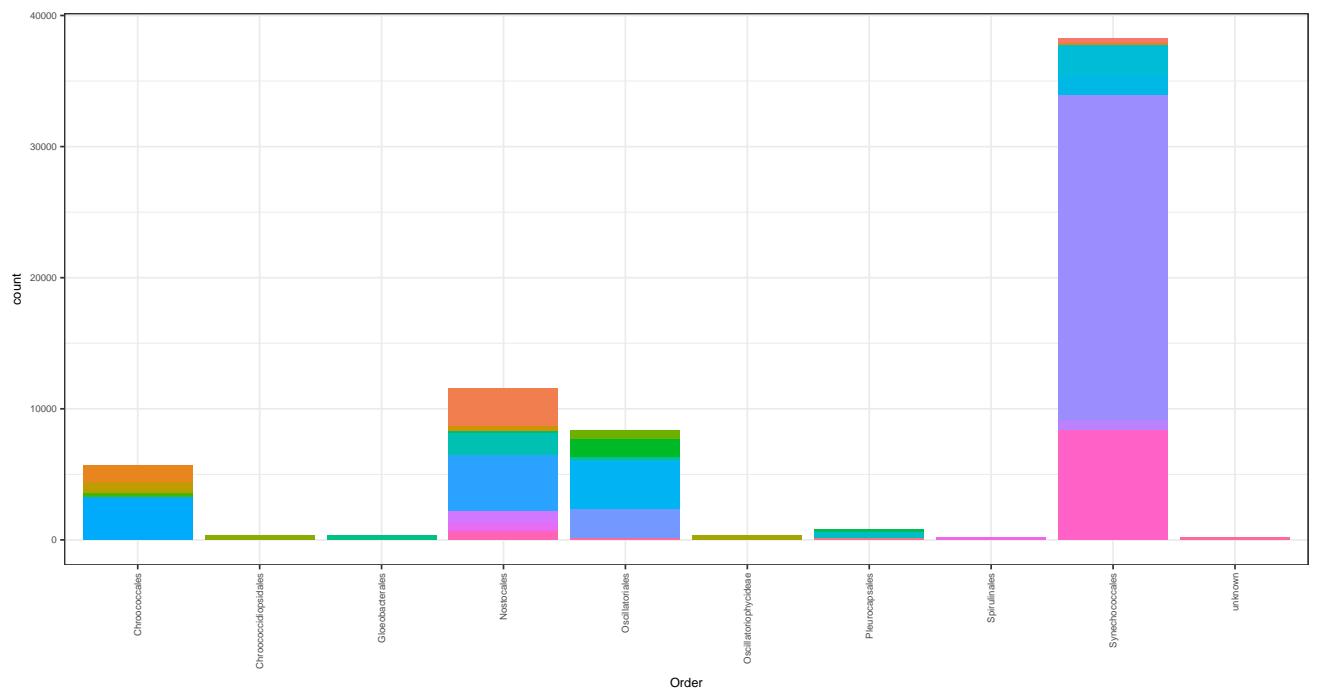


Figure 4.45: Diversity

Here is a reference to Recruitments after central pathways expansions coloured by taxa plot: Figure 4.45.

## Smash diversity

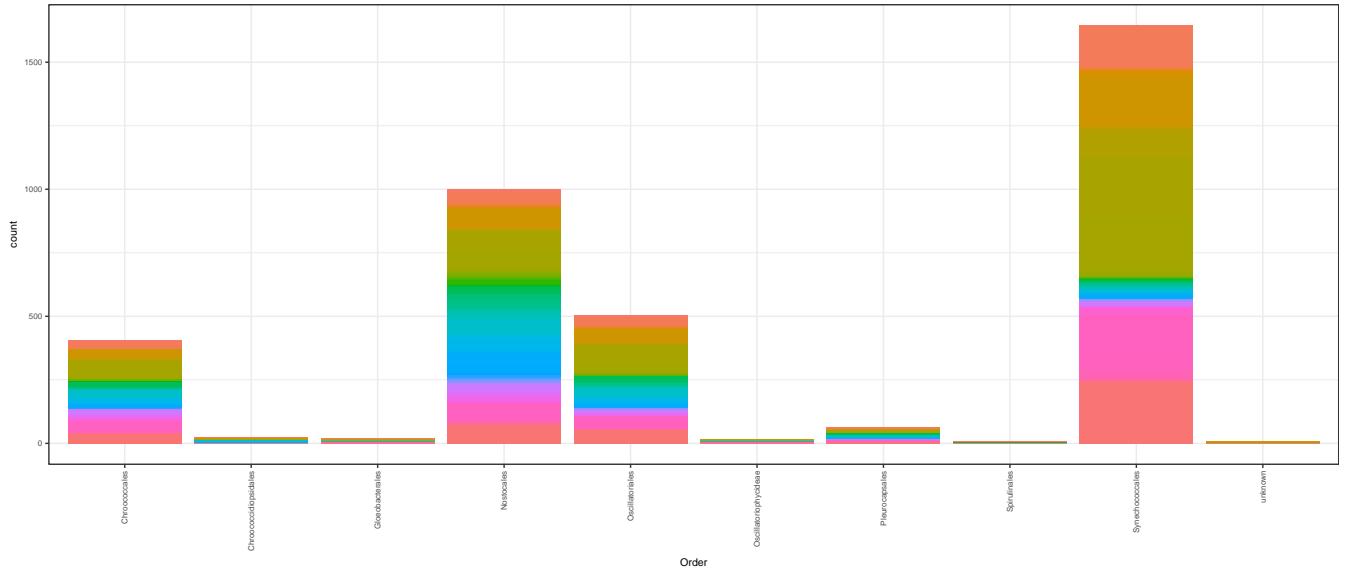


Figure 4.46: Smash

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

### 4.23.1 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

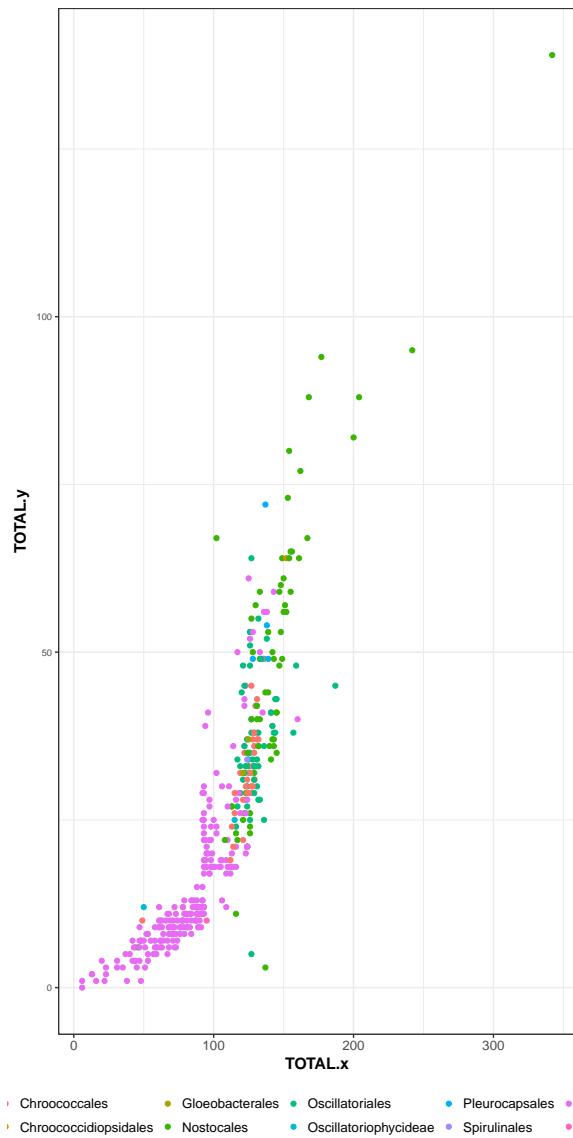


Figure 4.47: Correlation between central pathway expansions and anti-smash Natural products detection

Here is a reference to the expansions vs antismash NP's clusters plot: Figure 4.47.

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

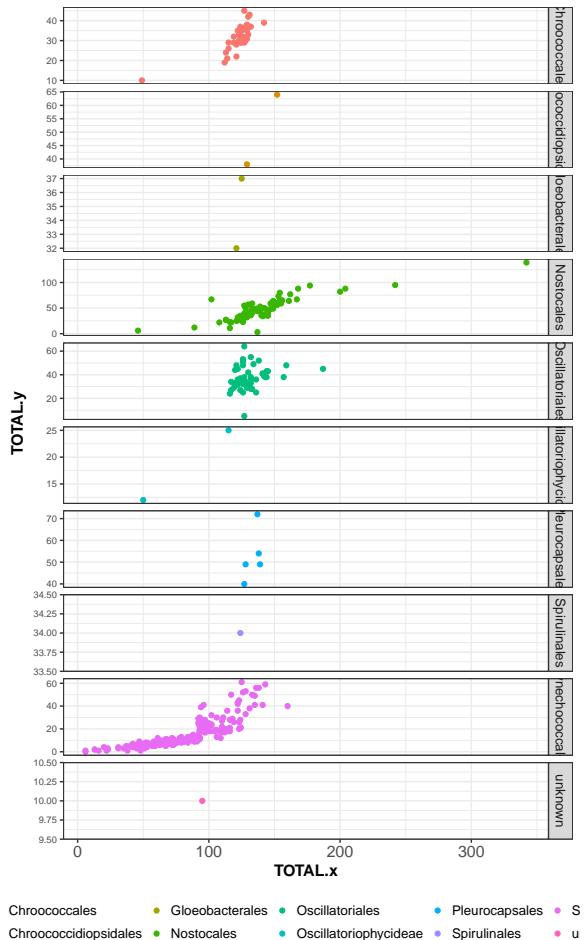


Figure 4.48: Correlation between central pathway expansions and anti-smash Natural products detection

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

## AntisMAsh vs Expansions by taxonomic Family

Natural products coloured by family

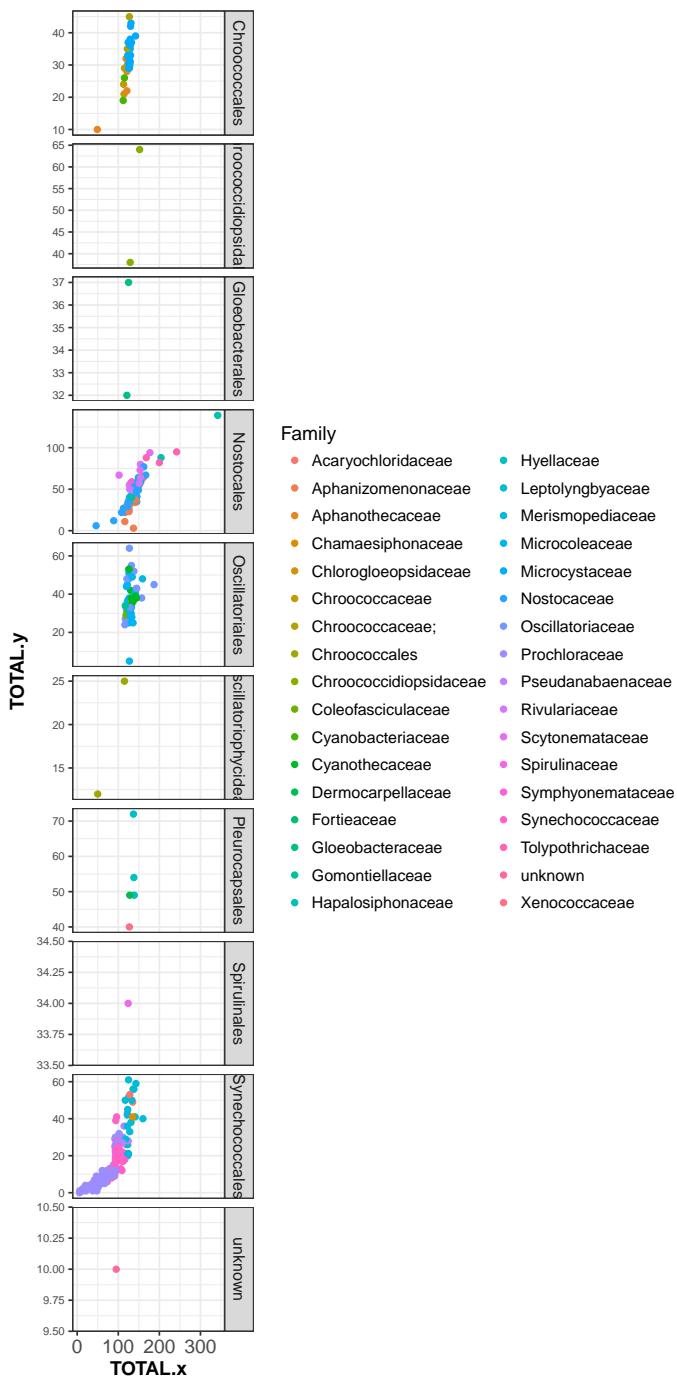


Figure 4.49: Natural products by family

Here is a reference to the Natural products coloured by family plot Figure 4.49.

## 4.24 Selected trees from EvoMining

Phosphoribosyl\_isomerase\_3 family

Figure from EvoMining

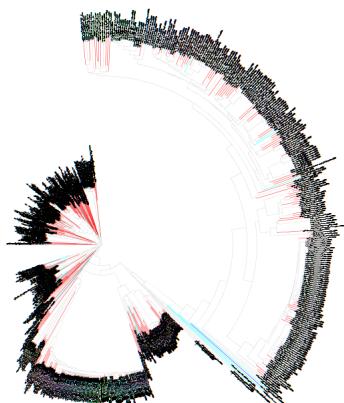


Figure 4.50: Phosphoribosyl isomerase EvoMiningtree

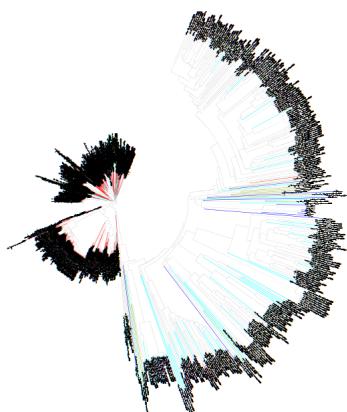


Figure 4.51: Phosphoglycerate dehydrogenase EvoMiningtree

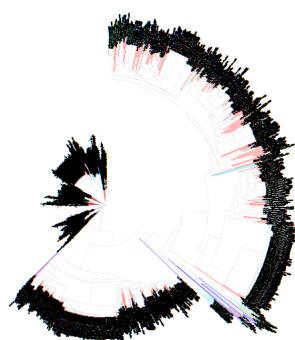


Figure 4.52: Phosphoserine aminotransferase EvoMiningtree

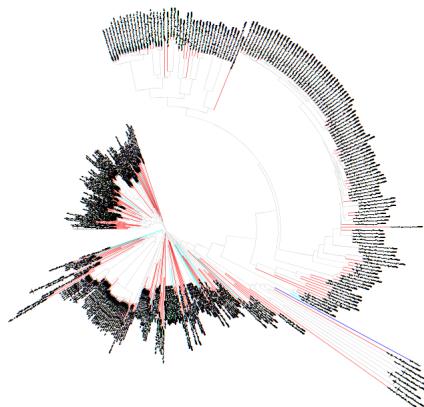


Figure 4.53: Triosephosphate isomerase EvoMiningtree

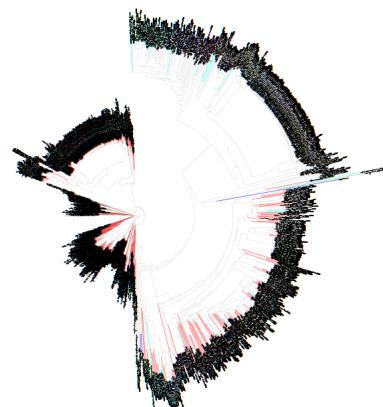


Figure 4.54: glyceraldehyde3phosphate dehydrogenase EvoMiningtree

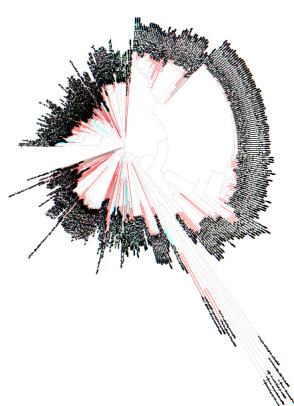


Figure 4.55: phosphoglycerate kinase EvoMiningtree

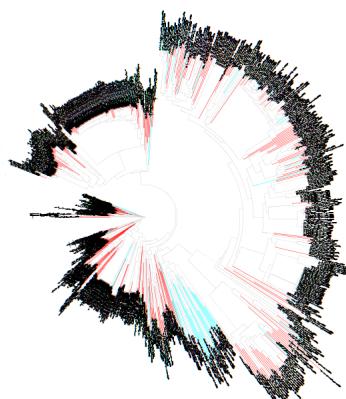


Figure 4.56: phosphoglycerate mutaseEvoMiningtree

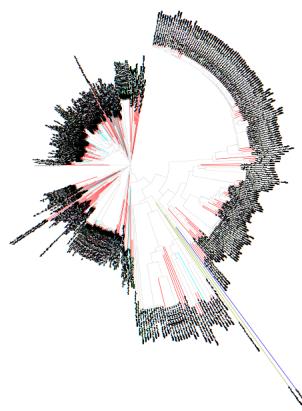


Figure 4.57: enolase EvoMiningtree

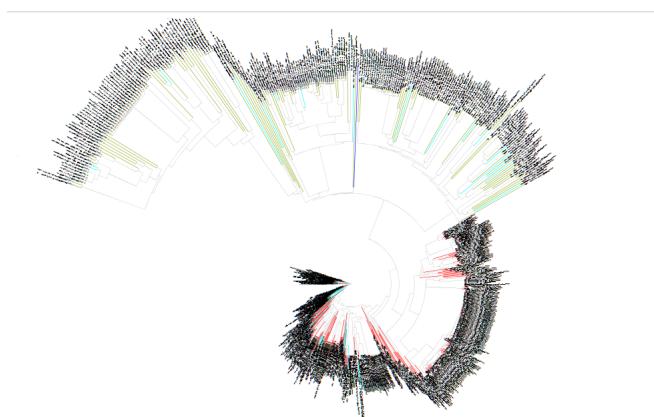


Figure 4.58: Pyruvate kinase EvoMiningtree

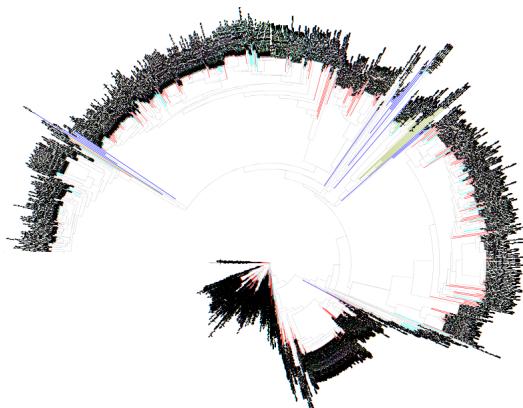


Figure 4.59: Aspartate transaminase EvoMiningtree

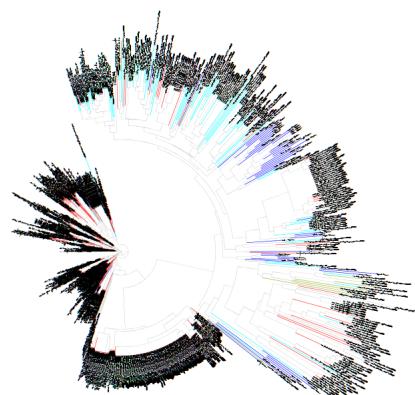


Figure 4.60: Asparagine synthase EvoMiningtree

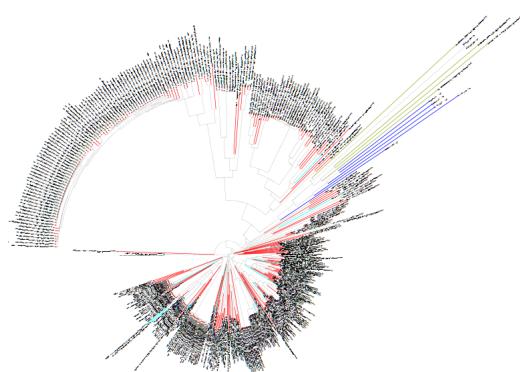


Figure 4.61: Aspartate kinase EvoMiningtree

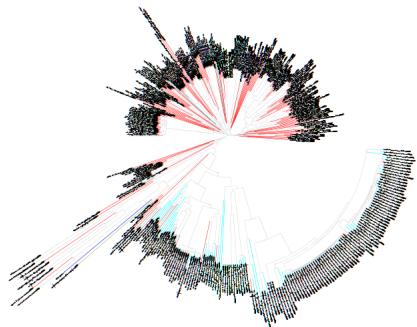


Figure 4.62: Aspartate semialdehyde dehydrogenase EvoMiningtree

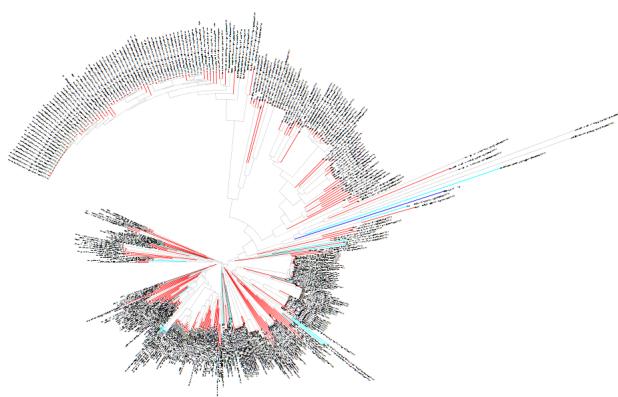


Figure 4.63: Homoserine dehydrogenase EvoMiningtree

# Chapter 5

## CORASON



# Chapter 6

## Desarrollo de CORASON como herramienta para organizar clusters biosintéticos y otras vecindades genómicas conservadas.

Figura [CorasonCaracteristicas] CORASON es una herramienta en línea de comandos para organizar filogenéticamente la variación de una familia de clusters. Dada la existencia de variantes de BGCs en bases de datos de linajes genómicos se diseñó CORASON. Corason se ejecuta en línea de comandos, su función es identificar el core génico de un BGC y utilizarlo para presentar una visualización de las variantes de la familia organizadas filogenéticamente

Los métodos de minado de genomas han acelerado el descubrimiento de nuevos clusters biosintéticos (BGCs). Casi cada nueva bacteria que es secuenciada aporta alguna novedad al pangenoma bacteriano conocido. Una fracción de estos genes novedosos formará parte de variantes de BGCs previamente conocidos aportando diversidad a las familias de clusters biosintéticos. La diversidad genética que existe en las familias de BGCs está directamente relacionada con cambios moleculares, incluso pequeñas variantes en un metabolito pueden ocasionar diferencias en su función biológica.

CORASON fue diseñado con las siguientes características: (i) una interfaz de línea de comando simple (ii) identificación del core génico del BGC (iii) la reconstrucción filogenética de la familia de BGCs utilizando la información del core (iv) salida visual en formato SVG, que muestra tanto la anotación funcional de los genes como la distancia respecto a sus ortólogos del cluster de referencia.

En este capítulo presento CORASON, CORe Analysis of Syntenic Orthologs to prioritize Natural products biosynthetic gene clusters, una herramienta para explorar la diversidad en el contenido de los BGCs así como su distribución en un linaje genómico proporcionado por el usuario.

## 6.1 Algoritmo y características de CORASON

Figura [CorasonPipeline] La herramienta corason localiza familias de clusters biosintéticos en un linaje genómico partiendo de un cluster y un gen de referencia. Todos los contextos genómicos que contengan ese gen y algún otro gen del cluster de referencia serán encontrados en el linaje seleccionado por el usuario. CORASON identifica el core génico de la familia. La información del core se utiliza para organizar filogenéticamente todos los miembros de la familia del BGC, es decir todas las variantes del BGC serán organizadas. El core génico está relacionado con el core de la molécula, la parte variable del BGC codifica enzimas accesorias que producen ornamentos i.e. variantes moleculares. Al cambiar de gen de referencia CORASON permite explorar otras familias de BGC que contengan las mismas modificaciones. Los resultados se presentan en una visualización que permite al mismo tiempo apreciar variación a nivel de presencia-ausencia de genes entre miembros de una familia de BGCs, como también apreciar variación a nivel de secuencia a través de un gradiente de color entre genes conservados entre una variante y el BGC de referencia.

CORASON permite la rápida identificación de variantes de BGCs y organiza los resultados utilizando una aproximación filogenética. En la figura [CorasonPipeline] se muestra esquemáticamente como dado un cluster de referencia y un gen localizado en el cluster, así como las secuencias genómicas de un linaje de referencia corason buscará todas las variantes del cluster de referencia que existen en el linaje genómico seleccionado.

## 6.2 Las familias de BGCs están formadas por variantes del BGC de referencia.

Así como existen familias génicas, un gen y todos sus homólogos, incluidos parálogos, ortólogos y xenólogos, también existen familias de BGCs. No es tan claro definir un BGC porque no tiene un codón de inicio y un codón de paro como un gen. En algunas ocasiones como en el caso de scytonemin todos los genes del BGC se expresan al recibir un estímulo, como los rayos UV en este caso. Otras veces en la producción del metabolito participan genes que no son necesariamente contiguos, y por ello al cambiar el BGC de organismo y realizar expresión heteróloga no se obtiene el mismo metabolito. Las fronteras es decir los genes de la orilla del BGC no siempre son claras.

Así pues, cuando se habla de un BGC no se debe pensar que este es igual en todos los organismos del linaje, es decir que tiene el mismo contenido génico. Hay variación tanto a nivel de contenido génico como a nivel de secuencia entre los genes en común. Esta variación produce la promiscuidad de producto, una familia de BGCs produce distintos productos a partir de los mismos precursores. En este trabajo vamos a tomar como BGC de referencia a los anotados en MIBiG, que son de los que se tienen datos

experimentales respecto al producto reportado. Todas sus variantes que contengan al menos dos genes en común, uno de referencia seleccionado por el usuario y otro cualquiera pero común con el BGC de referencia son considerados parte de la familia del BGC.

Como ejemplo de familias de BGCs podemos pensar a los operones. Un operón es el conjunto de genes dedicados a la síntesis de algún metabolito del metabolismo central, estos genes suelen encontrarse juntos en el cromosoma y transcribirse simultáneamente. Ejemplos de estos operones son los que producen los aminoácidos histidina y triptófano. Estos “BGCs” hace millones de años fueron posiblemente parte del metabolismo especializado y debido a su éxito se fijaron en lo que ahora vemos como BGCs conservados en ciertos linajes genómicos. En estos casos las fronteras son claras y la variación génica es poca. Aún así existen otros ejemplos donde puede constatarse variación en las rutas de síntesis de mecanismos centrales de metabolismo procariota. En el caso de histidina y triptófano, como hablaremos en el siguiente capítulo, es la variación a nivel de secuencia y no a nivel de composición génica la que produce diversidad de producto.

En oposición a los BGCs u operones de metabolismo conservado, están los BGCs del actual metabolismo especializado, así como se encuentran familias con mucha variación génica pueden encontrarse otras muy conservadas. En este capítulo presentaremos varios ejemplos de familias de BGCs.

## 6.3 Aplicaciones de CORASON en Actinobacteria y Pseudomonas

En el capítulo anterior las variantes del BGC scytonemin en Cyanobacteria fueron encontradas al aplicar CORASON a dicho linaje. En este capítulo presento ejemplos del uso de CORASON para investigar los patrones de conservación/variación en familias de BGCs en los linajes Actinobacteria y *Pseudomonas*. Primero, versiones iniciales de CORASON fueron usadas junto con cromatografía y espectrometría de masas (LC-MS) para estudiar la ecología y evolución de los sideróforos del tipo desferroxiaminas en Actinobacteria. Este estudio permitió la identificación de desG, un miembro de la familia penicilin amidasa responsable de la arilación de desferroxiaminas en actinomicetos acuáticos [REF].

En un segundo ejemplo, CORASON fue utilizado para investigar la existencia de variantes en Actinobacteria del cluster productor de arsenolípidos que se encuentra en *Streptomyces lividans*. En la tercera aplicación, se investigó la distribución en Actinobacteria de un BGC de *Streptomyces lividans* que es novedoso porque involucra una enzima que utiliza tRNA de la familia FemXAB. La predicción de producto de este BGC es un metabolito que contiene un grupo azoxy. Esta predicción ha sido confirmada utilizando LC-MS [Aguilar in prep].

Finalmente, un cuarto ejemplo es presentado: el de los contextos genómicos de *tauD* que codifica una dioxigenasa involucrada en el metabolismo de taurina en su papel de enzima del metabolismo conservado. En Actinobacteria *tauD* es parte de 15 BGCs reportados en MIBiG, aunque en *Pseudomonas* no existen BGCs reportados. EvoMining predice expansiones en este linaje genómico y CORASON predice conservación en los contextos genómicos de estas copias extra que guardan cierta similitud con variantes de los BGCs conocidos en Actinobacteria.

Estos cuatro ejemplos ilustran como CORASON puede ser utilizado tanto para priorizar nuevos BGCs como para ligar la variación génica de familias de BGCs con diversidad estructural. CORASON es una expansión de las habilidades de EvoMining [cruz-morales] de encontrar enzimas en el proceso de diversificación funcional. CORASON encuentra familias de BGCs, y presenta una rápida visualización de sus variantes. Ambas herramientas fueron desarrolladas para expandir nuestro conocimiento sobre nueva química mediante la genómica comparativa. CORASON está disponibles en su contenedor de docker en github <https://github.com/nselem/corason>.

## 6.4 Estas familias pueden clasificarse

Figura [CorasonSideróforos] Variantes del cluster biosintético de desferroxiamina fueron identificadas por CORASON en Actinobacterias de cuatro ciénegas. Una variante del BGC de desferroxiamina fue identificada. Esta variante se diferencia del cluster reportado por la ganancia de una penicilin amidasa. A este gen extra se le llamó *desG*. Fue verificado que la variación génica está ligada con la variación molecular

El hierro es necesario en el metabolismo de muchos seres vivos. Para poder utilizarlo las bacterias han desarrollado moléculas captadoras de hierro llamadas sideróforos. Un ejemplo de sideróforo es la molécula de desferroxiamina sintetizada por los genes *des* en Actinobacteria

## 6.5 Un BGC reportado puede tener muchas variantes que conforman una familia de BGCs

En el linaje de *Streptomyces* EvoMining obtuvo como predicción una arsenoenol piruvato sintasa en una rama del árbol de expansiones de la familia 3-fosfoshikimato-1-carboxivinyl transferasa (AroA). Estudios de mutagénesis y de expresión diferencial génica en presencia de arsénico confirmaron que en *Streptomyces coelicolor* y en *Streptomyces lividans* esta enzima pertenece a un BGC que sintetiza arsenolípidos [Pablo tesis]. El contexto genómico de AroA en estos organismos incluye una enzima PKS localizada a seis genes de distancia. AntiSMASH predice los BGCs tipo PKS de estas enzimas pero no incluye en ellos a la arsenoenol piruvato sintasa. El valor de EvoMining fue descubrir que esta copia extra de AroA estaba dedicada al metabolismo

especializado y por tanto bien podía tener su propio BGC o dada la cercanía con las PKS podía ser parte de estos PKS-BGCs. Esta ruta de síntesis de arseno compuestos, fue descubierta a partir de una rama de copias extra de AroA donde se apreciaba diversidad en las secuencias de aminoácidos. Una pregunta posible es si esta diversidad de secuencia a nivel de enzima ascendía a un siguiente nivel. Es decir si también existe diversidad a nivel del BGC, si se encuentran variantes con distintos patrones de presencia/ausencia en los genes que coponen al BGC de *S. coelicolor*. O más aun, quedaba por investigar si el BGC tenía cierto grado de conservación o era exclusivo de estos dos organismos *S. coelicolor* y *S. lividans*.

Una versión preliminar de CORASON, permitió visualizar las variantes de los contextos genómicos de las arsenol piruvato sintasas. Los genomas donde se buscaron estas variantes fueron seleccionados por una búsqueda de blast en NCBI en la base de datos no redundante como aparecía en 2016. Después de organizar manualmente los BGCs como se muestran a la izquierda de la [Fig Arseno-BGC], la visualización permitió distinguir 4 grupos de contextos conservados. El primero independiente de PKS y NRPS mostrado en morado, el segundo con una NRPS-PKS híbrida mostrado en un rectángulo verde y finalmente en rosa y naranja se muestran los grupos tercero y cuarto que contienen una PKS, los primeros del lado izquierdo y los segundos del lado derecho. Presumiblemente estos BGCs aunque contienen un core común producen distintos arseno compuestos.

[Fig Arseno-BGC] Contexto genómico conservado en las expansiones de AroA. La arsenoenol piruvato sintasa fue descubierta en un árbol de EvoMining como parte de una rama de expansiones de AroA en Actinobacteria. El contexto genómico de la arsenoenol piruvato sintasa de *S. coelicolor* tiene un core conservado en Actinobacteria. Este BGC está dedicado a la síntesis de metabolitos secundarios de tipo arsenolípidos. Primero se identificaron en otras secuencias genómicas contextos que contengan el homólogo de AroA y algún otro gen de su vecindad en *S. coelicolor*. A continuación, se ordenaron manualmente los contextos obtenidos y se pudo identificar al menos cuatro diferentes clases de BGCs. La primera clase mostrada en un rectángulo morado no contiene PKS o NRPS, la segunda enmarcada en verde contiene una PKS-NRPS híbrida. La tercera clase incluye una PKS a la izquierda y a más de cinco genes de distancia de la Arsenoenol piruvato sintasa y finalmente la última clase contiene una PKS a la derecha y a sólo un gen de distancia de esta enzima. Esta figura muestra que existen variantes de un BGC

La visualización realizada para los arsenolípidos no incluía ningún tipo de orden y era difícil distinguir grupos de BGCs. Por esta razón se pensó que el orden filogenético ya no de una enzima sino de el core del BGC ordenaría las variantes del BGC. Este orden mostraría en un continuo la dinámica genómica del BGC establecida por los procesos evolutivos. En el caso de los arsenolípidos se utilizó Orthocore para la identificación del core del BGC. En las siguientes versiones de CORASON esta característica fue implementada como parte del algoritmo.

## 6.6 CORASON analiza la conservación del contexto genómico de los EvoMining hits

La predicción de un BGC para la síntesis de un compuesto tipo azoxy es otro ejemplo de CORASON como complemento de EvoMining para la identificación de los dos niveles de promiscuidad, tanto a nivel familia enzimática como familia de BGCs. Este trabajo se encuentra en preparación para su publicación por Aguilar-Martínez.

Figura [CorasonAzoxy] Azoxy BGC

## 6.7 El contexto de una rama divergente de *tauD* está conservado en Pseudomonas

Figura[CorasonTauD] En Pseudomonas el gen *tauD* tiene un contexto que incluye genes de metabolismo especializado. Variantes de este contexto están distribuidas en varios organismos. En particular

## 6.8 CORASON se adaptó a la herramienta BiG-SCAPE de clasificación de familias de BGCs

BIG-SCAPE es una herramienta bioinformática para separar un conjunto de BGCs en familias de acuerdo al contenido, conservación y distribución de sus dominios. Un dominio es un conjunto de secuencia conservado de un gen, es considerado una unidad de las proteínas. Existe una base de datos de dominios llamada pFAM.

[CorasonBiGSCAPE] texto texto

CORASON complemento BIG-SCAPE al proporcionar un algoritmo para ordenar la diversidad dentro de la familia. A la vez CORASON permite conectar mediante la evolución a familias aparentemente separadas por BiGSCAPE.

## 6.9 BiG SCAPE y CORASON identificaron nuevos productos variantes de la familia de BGCs Rimosamide - Detoxin en Actinobacteria

En esta sección utilizamos CORASON y BiG-SCAPE para analizar una base de datos de miles de genomas de Actinobacteria. Con estas herramientas se organizó la diversidad biosintética de las familias de BGC detoxin y rimosamide [REF]. El análisis reveló diversidad tanto en los géneros de los organismos que contienen a esta familia, y en la composición génica del BGC. Entre los géneros con alguna variante del BGC están Amycolatopsis, Streptomyces. El core conservado de los BGCs detoxin y rimosamide está compuesto por una NRPS, una NRPS/PKS híbrida, y un homólogo de tauD. Este homólogo fue sugerido por un análisis de EvoMining8, En E.coli tauD se encuentra en el operón tauABCD [MM1] La ruta de síntesis de rimosamide difiere de la de detoxins porque tiene una NRPS adicional, que codifica para una modificación del core de molécula detoxin/rimosamide con isobutyrate y glycine39.

Figura [TauDActinobacteria] texto

El hecho de que el gen *tauD* estuviera presente en todos los miembros de la familia captó nuestra atención. El producto del gen *tauD* pertenece a la super familia de enzimas Fe(II)/ $\alpha$ -ketoglutarato-dependientes hidroxilasas. En particular *tauD* codifica una  $\alpha$ -ketoglutarato-dependiente taurina dioxigenasa involucrada en la asimilación de sulfito por la liberación oxigenólica del aminoácido taurina41. Interesantemente, esta familia también incluye enzimas en linajes como hongos, bacterias y plantas. Dichas enzimas catalizan hydroxylations, desaturations, ring expansions and ring formations, among other chemical transformations. A la fecha, el rol de TauD en la biosíntesis de los metabolitos detoxin y rimosamide aún es desconocido, se ha sugerido que es responsable de la oxidación de la prolina observada en algunos análogos39.

Para identificar variantes de los BGCs relacionados a detoxin y rimosamide dentro de la rama de metabolismo especializado los 1175 BGCs del dataset que contenían un homólogo de *tauD* se pasaron por un análisis combinado de BiG-SCAPE/CORASON. Los BGCs fueron procesados con CORASON usando *tauD* como query gene. Teniendo como resultado que es el único miembro del ‘BGC core’. Es importante notar que el core del BGC puede contener entre 1 y 3 genes, [La NRPS, y la NRPS-PKS híbrida quedan en este ejemplo fuera del core debido a gaps en la secuencia del genoma de organismos como *Streptomyces humi*, *Streptomyces Spectabilis* y *Amycolatopsis Vancorresmycina*]



## Chapter 7

**CORASON sugiere que las familias detoxin y rimosamide pertenecen a un amplio clan de familias dedicadas a la síntesis de péptidos.**

El análisis de CORASON reveló que las familias de los BGCs detoxin y rimosamide GCFs identificados en BiG-SCAPE eran parte de todo un clan de biosíntesis de péptidos que incluía clados inexplorados [MM10] del phylum Actinobacteria [Fig Tau-DActinobacteria]. La organización filogenética de los BGCs provista por CORASON, reveló familias de BGCs que fueron omitidas debido al umbral utilizado por el algoritmo de clustering de BiG-SCAPE. Esto debido a la cercanía de genes biosintéticos detectados por antiSMASH como suficientemente diferentes como para ser clasificados en otras familias (ver Fig. S10).

Hipotetizamos que las toxins codificadas por los BGCs en los clados inexplorados contendrán cierta novedad química relacionada con las variaciones genéticas. Afortunadamente, 40 de las 152 cepas identificadas como portadoras de un BGCs estaban representadas en nuestros datos metabolómicos de 363–cepas por LC-MS/MS. Análisis de redes moleculares de estos datos indicaron la presencia de tres toxin conocidas, cuatro rimosamide conocidas y otras 103 variantes de detoxin o rimosamide (Fig. 3c), el basto universo químico sugerido por el análisis BiG-SCAPE/CORASON.

## 7.1 Clados selectos del árbol de CORASON que contiene a las familias rimosamide/detoxin contienen diversidad génica que correlaciona con novedad química.

Tres de los clados que codifican para detoxin BGC fueron identificados por BiG-SCAPE dentro del árbol de CORASON capturaron nuestro interés ([Fig TauDActinobacteria], in colored boxes). En esta sección se describe el trabajo experimental realizado por Michael Mulloney del grupo de colaboradores de West Chicago, de los tres clados y específicamente de los organismos que seleccioné como candidatos a presentar diversidad química. ### El clado P450/enoyl agrega una heptanamida al core molecular detoxin/rimosamide El primer clado es el ‘P450/enoyl clade’ contiene genes como el citocromo P450 y un enoyl-CoA hidratasa/isomerasa dentro de cada uno de sus BGCs. Este clado está marcado en rojo [Fig TauDActinobacteria]. Análisis de datos por tandem MS de extractos de *Streptomyces* sp. NRRL S-325, que se encuentra dentro de este clado, llevó al descubrimiento de la detoxin S1 (1; Fig TauDActinobacteria, S16–17). Este nuevo análogo contiene una cadena lateral de heptanamide, una extructura única entre las detoxins y rimosamides cuya instalación posiblemente depende de la enzima enoyl-CoA hidratase/isomerasa.

### 7.1.1 El superclado spectinomycin/ detoxin-rimosamide-clan produce cinco variantes de detoxin.

El segundo clado de interés,fue nombrado el ‘supercluster clade’ (Fig. 5, en verde claro), comprende los BGCs con genes de detoxin adjacentes al cluster que produce spectinomycin42 [Fig TauDActinobacteria]. El cluster de spectinomycin (MIBiG BGC0000715) contiene en su periferia al gen tauD como se muestra en la línea gris punteada de la [Fig TauDActinobacteria]. La secuencia del cluster de spectinomycin depositada en MIBiG es la única secuencia disponible de *Streptomyces spectabilis* NRRL 2792. Como no se sabe que tauD participe en la síntesis de spectinomycin se hipotetizó que pueden existir los genes del cluster de detoxin al lado de los genes del BGC de spectinomycin en *S. spectabilis* NRRL 2792. Adquirimos esta cepa para determinar si el análisis de CORASON podía ayudar a la predicción de detoxin basado solamente en la presencia del query gen pero en ausencia completa de la secuencia del BGC de detoxin. Análisis en tandem de espectrometría de masas de extracto *S. spectabilis* NRRL 2792 reveló la producción de cinco compuestos tipo detoxin, incluyendo detoxin N1 (2; Fig TauDActinobacteria, S15), detoxin N2 (3; Fig. S15) y su análogo acetoxylated, detoxin N3 (4; Fig. S15).

Los tiempos de retención de ionies y los patrones de fragmentación de los últimos dos compuestos también fueron observados en extractos de *Streptomyces* sp. NRRL B-1347 parte del clado del supercluster, confirmando la habilidad de CORASON para

guiar un descubrimiento, mediante la utilización de la filogenia a pesar de lo limitado de los datos en la cepa NRRL-2792. Análisis de LC-MS de cultivos de NRRL-2792 suplementados con isotopos estables etiquetados de amino ácidos corroboraron las predicciones estructurales basadas en los análisis de la cepa cercana *Streptomyces* sp. NRRL B-1347 (Fig TauDActinobacteria, S26–31). Los tres nuevos análogos incorporan completamente el <sup>13</sup>C6-isoleucine, pero d7-proline solo es incorporado en el compuesto 3. La pérdida de un deuterio de d7-proline en 2 y 4 soporta la asignación de acetoxylation del anillo de pyrrolidine común en detoxins y rimosamides<sup>39,43–45</sup>. Características especiales únicas a la serie de N-detoxins incluyen la incorporación de una tirosina N-formylated en 3 y 4 en lugar de fenilalanina, el residuo típico de detoxin/rimosamide, lo que es soportado por la incorporación del anillo d4-tyrosine. El compuesto 2 la incorporación única de un residuo derivado del triptófano en esta posición, haciendo evidente la retención de cuatro deuterios cuando en experimentos de alimentación de indole-d5-tryptophan (Fig. S27). Aunque los datos de MS fueron insuficientes para desenmascarar esta estructura, el compuesto 2 fue producido por *S. spectabilis* NRRL 2792 en suficiente abundancia para el aislamiento y la elucidación estructural por NMR. Varios experimentos 1D y 2D confirmaron las asignaciones por datos de MS y establecieron una N-acetylated kynurenine como la estructura derivada de triptófano en 2 (Figs. S18–S27).

### 7.1.2 El clado *Amycolatopsis P450* produce cinco variantes de toxin.

El tercer clado que se estudió del super clan al que pertenecen las familias rimosamide-detoxin, contiene BGCs casi enteramente provenientes del género *Amycolatopsis*. Este clado está marcado en morado en la [Fig TauDActinobacteria]. Este clado de BGCs también contienen un gen P450 único entre los BGCs del árbol, así que fue llamado el clado ‘*Amycolatopsis/P450* clade’. Aunque no se contaba con datos metabolómicos de las cepas del clado de BGCs definidos por BiG-SCAPE como una gen cluster family (GCF), la visualización filogenética de CORASON permitió la selección de una cepa de *Amycolatopsis* de la que se tenían datos metabolómicos con un BGC muy similar, y que también contiene el gen P450 ([Fig TauDActinobacteria], línea gris cerca del clado *Amycolatopsis/P450*). Análisis de datos de tandem MS de extracto fermentado de *Amycolatopsis jejuensis* NRRL B-24427 reveló isómeros de toxins P1 (5; [Fig TauDActinobacteria], S15) que contienen tirosina, P2 (6; [Fig TauDActinobacteria], S15) mostrando fenilalanina y una valina hidroxilada, así como la toxin P3, un análogo cercano libre de hidroxilación (7; [Fig TauDActinobacteria], S15). Una validación de la asignación de aminoácidos observados en los patrones de fragmentación de MS/MS se consiguió mediante el uso de experimentos de incorporación de isotopos estables etiquetados de aminoácidos (Figs. S33–S36, S38–S41, and S43–S44). ## CORASON permitió expandir las capacidades de EvoMining y explorar la promiscuidad de familias de BGCs de enzimas divergentes de metabolismo central. Nuestros resultados ilustran como BiG-SCAPE puede identificar conjuntos de BGCs relacionados, en un gran

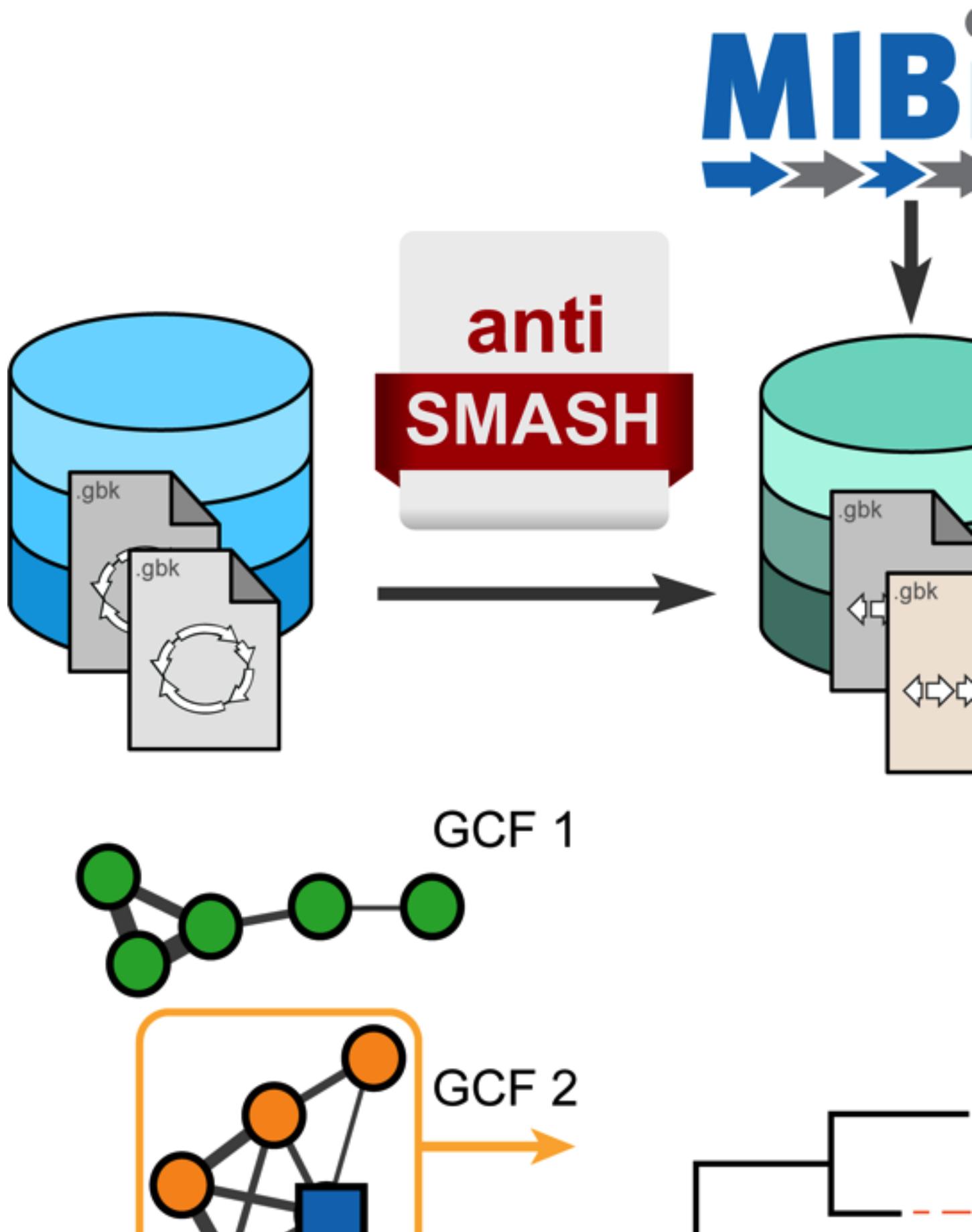
número de secuencias de genomas. Además al usar CORASON para reconstruir las filogenias de BGC para ordenar visualmente la evolución de un cluster biosintético y su diversidad proveen herramientas poderosas para el descubrimiento de nuevos clados de BGC que codifican en consecuencia para nueva química. Respecto a los BGCs detoxin/rimosamide, CORASON mostró habilidad para ayudar a minar bases de datos genómicas y descubrir siete nuevas detoxins. Específicamente, la organización de las variantes de los BGC facilitó la identificación de los correspondientes variaciones en la estructura química -la presencia de un enoyl-CoA hidratasa/isomerasa corresponde a la familia de amida ácido graso detoxin S1 y la presencia de un gen P450 corresponde a la presencia de hidroxilaciones en detoxins P1–P3.

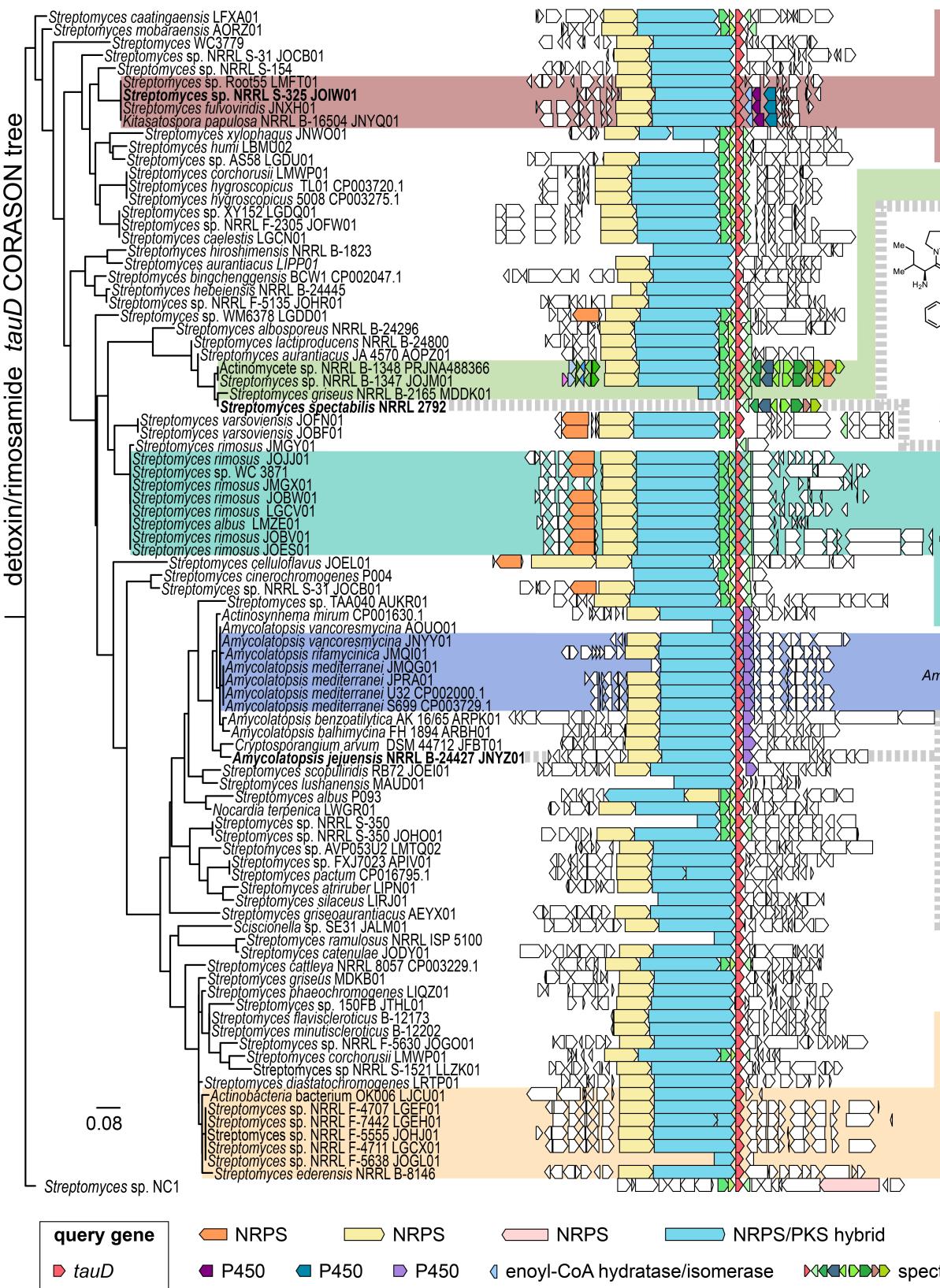
# **Chapter 8**

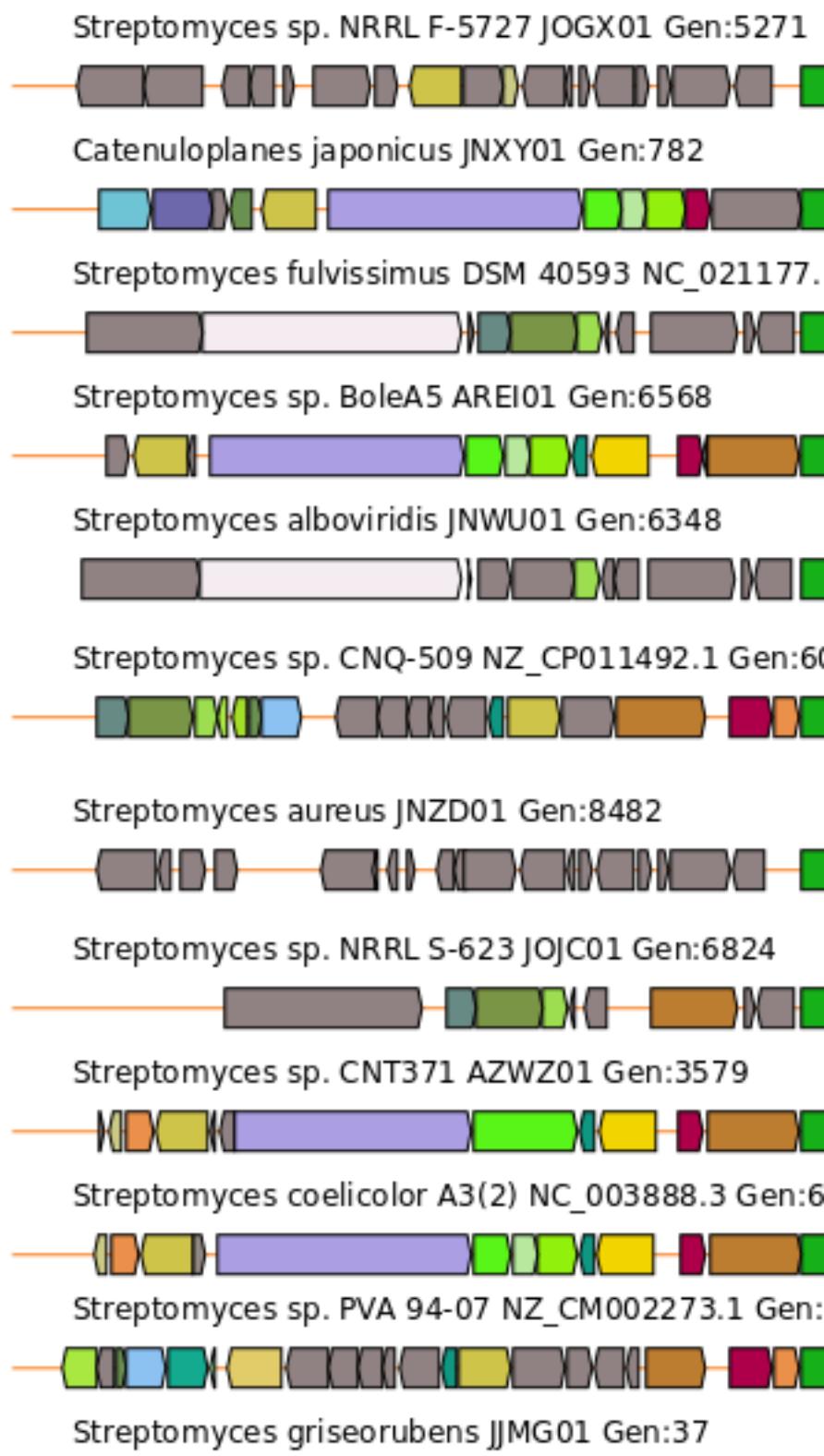
## **Discusión**

El cluster des está muy conservado, no así los de arsenolípidos, ni los de rimosamide

```
{r chapter4, child = 'chap4.Rmd'}
```







# 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

## 4.1 Discussion

Finalmente, una pregunta abierta es aún dada una enzima promiscua es la población promiscua, y son sólo unos confórmeros o todas y cada una de las enzimas son promiscuas

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") ## this shoul alwa
```

```
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)
```

# Appendix B

## The Second Appendix, Open source code on this document

### B.1 R markdown

Thanks to Rmакdown 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/>

```
Docker antiSMASH run_antismash /.gbk ~/as_results --knownclusterblast --inclusive  
--cf_threshold 0.7
```



# **Appendix C**

## **The third Appendix, Other contributions during my phd**

### **C.1 Accepted**

-Evomining identifies Arsenolipids biosynthetic cluster

### **C.2 Submitted**

- Siderophore micrococcus cluster identified by CORASON
- CORASON find genomes that owns cluster from cyanobacterial metagenome
- Streptomyces central pathways expansions

### **C.3 On preparation**

- PriA non Darwinian trayectories  
poner figura James unidades docking



# References

1. Khersonsky O, Tawfik DS. Enzyme promiscuity: A mechanistic and evolutionary perspective. *Annual Review of Biochemistry*. 2010;79:471–505.
2. Copley SD. Enzymes with extra talents: Moonlighting functions and catalytic promiscuity. *Current Opinion in Chemical Biology* [Internet]. 2003 Apr [cited 2017 Feb 8];7(2):265–72. Available from: <https://www.sciencedirect.com/science/article/pii/S1367593103000322>
3. Hult K, Berglund P. Enzyme promiscuity: Mechanism and applications. *Trends in Biotechnology* [Internet]. 2007 May [cited 2017 Feb 8];25(5):231–8. Available from: <https://www.sciencedirect.com/science/article/pii/S016777990700073X>
4. O'Brien PJ, Herschlag D. Catalytic promiscuity and the evolution of new enzymatic activities. *Chemistry & Biology* [Internet]. 1999 Apr [cited 2017 Feb 9];6(4):R91–R105. Available from: <http://www.sciencedirect.com/science/article/pii/S1074552199800337>
5. Barona Gómez F, Hodgson DA. Occurrence of a putative ancient like isomerase involved in histidine and tryptophan biosynthesis. *EMBO reports* [Internet]. 2003 Mar [cited 2017 Feb 8];4(3):296–300. Available from: <http://embor.embopress.org/content/4/3/296>
6. 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* [Internet]. 2014 Jun [cited 2017 Feb 9];82(6):887–96. Available from: <http://onlinelibrary.wiley.com/doi/10.1002/prot.24575/abstract>
7. 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* [Internet]. 2007 Jan [cited 2017 Feb 8];21(1):368–72. Available from: <http://dx.doi.org/10.1021/ef0602168>
8. Li C, Henry CS, Jankowski MD, Ionita JA, Hatzimanikatis V, Broadbelt LJ. Computational discovery of biochemical routes to specialty chemicals. *Chemical Engineering Science* [Internet]. 2004 Nov [cited 2017 Feb 8];59(22–23):5051–

60. Available from: <https://www.sciencedirect.com/science/article/pii/S0009250904006669>
9. Glasner ME, Gerlt JA, Babbitt PC. Evolution of enzyme superfamilies. *Current Opinion in Chemical Biology* [Internet]. 2006 Oct [cited 2017 Feb 9];10(5):492–7. Available from: <https://www.sciencedirect.com/science/article/pii/S1367593106001177>
10. Baier F, Copp JN, Tokuriki N. Evolution of Enzyme Superfamilies: Comprehensive Exploration of Sequence–Function Relationships. *Biochemistry* [Internet]. 2016 Nov [cited 2017 Feb 8];55(46):6375–88. Available from: <http://dx.doi.org/10.1021/acs.biochem.6b00723>
11. Bloom JD, Romero PA, Lu Z, Arnold FH. Neutral genetic drift can alter promiscuous protein functions, potentially aiding functional evolution. *Biology Direct* [Internet]. 2007 [cited 2017 Feb 8];2:17. Available from: <http://dx.doi.org/10.1186/1745-6150-2-17>
12. Nath A, Atkins WM. A Quantitative Index of Substrate Promiscuity. *Biochemistry* [Internet]. 2008 Jan [cited 2017 Feb 9];47(1):157–66. Available from: <http://dx.doi.org/10.1021/bi701448p>
13. Zou T, Rissó 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* [Internet]. 2015 Jan [cited 2017 Feb 9];32(1):132–43. Available from: <https://academic.oup.com/mbe/article/32/1/132/2925568/Evolution-of-Conformational-Dynamics-Determines>
14. Firn RD, Jones CG. A Darwinian view of metabolism: Molecular properties determine fitness. *Journal of Experimental Botany* [Internet]. 2009 Mar [cited 2017 Feb 8];60(3):719–26. Available from: <https://academic.oup.com/jxb/article/60/3/719/452667/A-Darwinian-view-of-metabolism-molecular>
15. Jia B, Cheong G-W, Zhang S. Multifunctional enzymes in archaea: Promiscuity and moonlight. *Extremophiles* [Internet]. 2013 Mar [cited 2017 Feb 8];17(2):193–203. Available from: <http://link.springer.com/article/10.1007/s00792-012-0509-1>
16. Aharoni A, Gaidukov L, Khersonsky O, Gould SM, Roodveldt C, Tawfik DS. The 'evolvability' of promiscuous protein functions. *Nature Genetics* [Internet]. 2005 Jan [cited 2017 Feb 8];37(1):73–6. Available from: <http://www.nature.com/ng/journal/v37/n1/full/ng1482.html>
17. Jensen. Enzyme Recruitment in Evolution of New Function. *Annual Review of Microbiology* [Internet]. 1976 [cited 2017 Feb 8];30(1):409–25. Available from: <http://dx.doi.org/10.1146/annurev.mi.30.100176.002205>
18. Pandya C, Farelli JD, Dunaway-Mariano D, Allen KN. Enzyme Promiscuity:

- Engine of Evolutionary Innovation. *Journal of Biological Chemistry* [Internet]. 2014 Oct [cited 2017 Feb 9];289(44):30229–36. Available from: <http://www.jbc.org/content/289/44/30229>
19. Dean AM, Thornton JW. Mechanistic approaches to the study of evolution. *Nature reviews Genetics* [Internet]. 2007 Sep [cited 2017 Feb 8];8(9):675–88. Available from: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2488205/>
  20. Nobeli I, Favia AD, Thornton JM. Protein promiscuity and its implications for biotechnology. *Nature Biotechnology* [Internet]. 2009 Feb [cited 2017 Feb 9];27(2):157–67. Available from: <http://www.nature.com/nbt/journal/v27/n2/full/nbt1519.html>
  21. Hopkins AL. Drug discovery: Predicting promiscuity. *Nature* [Internet]. 2009 Nov [cited 2017 Feb 8];462(7270):167–8. Available from: <http://www.nature.com/nature/journal/v462/n7270/full/462167a.html>
  22. 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* [Internet]. 2010 Dec [cited 2017 Feb 9];38(12):2195–203. Available from: <http://dmd.aspetjournals.org/content/38/12/2195>
  23. Eichborn J von, Murgueitio MS, Dunkel M, Koerner S, Bourne PE, Preissner R. PROMISCUOUS: A database for network-based drug-repositioning. *Nucleic Acids Research* [Internet]. 2011 Jan [cited 2017 Feb 9];39(suppl\_1):D1060–6. Available from: [https://academic.oup.com/nar/article/39/suppl\\_1/D1060/2506056/PROMISCUOUS-a-database-for-network-based-drug](https://academic.oup.com/nar/article/39/suppl_1/D1060/2506056/PROMISCUOUS-a-database-for-network-based-drug)
  24. Zhang W, Dourado DFAR, Fernandes PA, Ramos MJ, Mannervik B. Multi-dimensional epistasis and fitness landscapes in enzyme evolution. *Biochemical Journal* [Internet]. 2012 Jul [cited 2017 Feb 9];445(1):39–46. Available from: <http://www.biochemj.org/content/445/1/39>
  25. Sanchez-Ruiz JM. On promiscuity, changing environments and the possibility of replaying the molecular tape of life. *Biochemical Journal* [Internet]. 2012 Jul [cited 2017 Feb 9];445(1):e1–3. Available from: <http://www.biochemj.org/content/445/1/e1>
  26. 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* [Internet]. 2015 Sep [cited 2017 Feb 8];83(9):1625–31. Available from: <http://onlinelibrary.wiley.com/doi/10.1002/prot.24847/abstract>
  27. Patrick WM, Quandt EM, Swartzlander DB, Matsumura I. Multicopy Suppression Underpins Metabolic Evolvability. *Molecular Biology and Evolution* [Internet]. 2007 Dec [cited 2017 Feb 9];24(12):2716–22. Available from: <https://academic.oup.com/mbe/article/24/12/2716/978039/Multicopy-Suppression-Underpins-Metabolic-Evolvability>

- Suppression-Underpins-Metabolic
28. Notebaart RA, Szappanos B, Kintses B, Pál F, Györkei Á, Bogos B, et al. Network-level architecture and the evolutionary potential of underground metabolism. *Proceedings of the National Academy of Sciences* [Internet]. 2014 Aug [cited 2017 Feb 9];111(32):11762–7. Available from: <http://www.pnas.org/content/111/32/11762>
  29. Linster CL, Van Schaftingen E, Hanson AD. Metabolite damage and its repair or pre-emption. *Nature Chemical Biology* [Internet]. 2013 Feb [cited 2017 Feb 8];9(2):72–80. Available from: <http://www.nature.com/nchembio/journal/v9/n2/full/nchembio.1141.html>
  30. 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* [Internet]. 2015 Jan [cited 2017 Feb 8];32(1):100–8. Available from: <https://academic.oup.com/mbe/article/32/1/100/2925554/Differential-Effects-of-a-Mutation-on-the-Normal>
  31. 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* [Internet]. 2013 Jun [cited 2017 Feb 8];20(6):796–805. Available from: <https://www.sciencedirect.com/science/article/pii/S1074552113001701>
  32. 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* [Internet]. 2014 Jul [cited 2017 Jan 31];5(4). Available from: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4161248/>
  33. Soskine M, Tawfik DS. Mutational effects and the evolution of new protein functions. *Nature Reviews Genetics* [Internet]. 2010 Aug [cited 2017 Feb 9];11(8):572–82. Available from: <http://www.nature.com/nrg/journal/v11/n8/full/nrg2808.html>
  34. Halachev MR, Loman NJ, Pallen MJ. Calculating Orthologs in Bacteria and Archaea: A Divide and Conquer Approach. *PLOS ONE* [Internet]. 2011 Dec [cited 2016 Sep 16];6(12):e28388. Available from: <http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0028388>
  35. Kislyuk AO, Haegeman B, Bergman NH, Weitz JS. Genomic fluidity: An integrative view of gene diversity within microbial populations. *BMC Genomics* [Internet]. 2011 [cited 2017 Jan 26];12:32. Available from: <http://dx.doi.org/10.1186/1471-2164-12-32>
  36. Carbonell P, Faulon J-L. Molecular signatures-based prediction of enzyme promiscuity. *Bioinformatics* [Internet]. 2010 Aug [cited 2017 Feb 8];26(16):2012–9. Avail-

- able from: <https://academic.oup.com/bioinformatics/article/26/16/2012/215921/Molecular-signatures-based-prediction-of-enzyme>
37. 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 [Internet]. 2012 Jun [cited 2017 Feb 8];7(6). Available from: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3377604/>
38. Nagao C, Nagano N, Mizuguchi K. Prediction of Detailed Enzyme Functions and Identification of Specificity Determining Residues by Random Forests. PLOS ONE [Internet]. 2014 Jan [cited 2017 Feb 8];9(1):e84623. Available from: <http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0084623>
39. 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 [Internet]. 2015 Jun [cited 2017 Jan 31];15. Available from: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4462073/>
40. 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 [Internet]. 2012 Jun [cited 2017 Feb 8];8(6):e1002558. Available from: <http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1002558>
41. 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 [Internet]. 2013 Sep [cited 2017 Jan 31];30(9):2024–34. Available from: <https://academic.oup.com/mbe/article/30/9/2024/1000280/Evolution-of-Substrate-Specificity-in-a-Recipient>
42. Verdel-Aranda K, López-Cortina ST, Hodgson DA, Barona-Gómez F. Molecular annotation of ketol-acid reductoisomerases from Streptomyces reveals a novel amino acid biosynthesis interlock mediated by enzyme promiscuity. Microbial Biotechnology [Internet]. 2015 Mar [cited 2017 Feb 9];8(2):239–52. Available from: <http://onlinelibrary.wiley.com/doi/10.1111/1751-7915.12175/abstract>
43. Nesvizhskii AI, Vitek O, Aebersold R. Analysis and validation of proteomic data generated by tandem mass spectrometry. Nature Methods [Internet]. 2007 Oct [cited 2017 Feb 9];4(10):787–97. Available from: <http://www.nature.com/nmeth/journal/v4/n10/abs/nmeth1088.html>
44. Medema MH, Fischbach MA. Computational approaches to natural product discovery. Nature Chemical Biology [Internet]. 2015 Sep [cited 2017 Jan 24];11(9):639–48. Available from: <http://www.nature.com/nchembio/journal/v11/n9/full/>

- nchembio.1884.html
45. Campbell I. Biophysical Techniques - Paperback - Iain D. Campbell - Oxford University Press. 2012.
  46. Yang JY, Sanchez LM, Rath CM, Liu X, Boudreau PD, Bruns N, et al. Molecular Networking as a Dereplication Strategy. *Journal of Natural Products* [Internet]. 2013 Sep [cited 2017 Feb 9];76(9):1686–99. Available from: <http://dx.doi.org/10.1021/np400413s>
  47. Köcher T, Superti-Furga G. Mass spectrometry-based functional proteomics: From molecular machines to protein networks. *Nature Methods* [Internet]. 2007 Oct [cited 2017 Feb 10];4(10):807–15. Available from: <http://www.nature.com/nmeth/journal/v4/n10/full/nmeth1093.html>
  48. James LC, Tawfik DS. Conformational diversity and protein evolution – a 60-year-old hypothesis revisited. *Trends in Biochemical Sciences* [Internet]. 2003 Jul [cited 2017 Feb 8];28(7):361–8. Available from: <https://www.sciencedirect.com/science/article/pii/S096800040300135X>
  49. Parisi G, Zea DJ, Monzon AM, Marino-Buslje C. Conformational diversity and the emergence of sequence signatures during evolution. *Current Opinion in Structural Biology* [Internet]. 2015 Jun [cited 2017 Feb 9];32:58–65. Available from: <https://www.sciencedirect.com/science/article/pii/S0959440X15000147>
  50. Javier Zea D, Miguel Monzon A, Fornasari MS, Marino-Buslje C, Parisi G. Protein Conformational Diversity Correlates with Evolutionary Rate. *Molecular Biology and Evolution* [Internet]. 2013 Jul [cited 2017 Feb 8];30(7):1500–3. Available from: <https://academic.oup.com/mbe/article/30/7/1500/972515/Protein-Conformational-Diversity-Correlates-with>
  51. Gatti-Lafranconi P, Hollfelder F. Flexibility and Reactivity in Promiscuous Enzymes. *ChemBioChem* [Internet]. 2013 Feb [cited 2017 Feb 8];14(3):285–92. Available from: <http://onlinelibrary.wiley.com/doi/10.1002/cbic.201200628/abstract>
  52. 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 Streptomycetes. *Genome Biology and Evolution* [Internet]. 2016 Jun [cited 2017 Jan 24];8(6):1906–16. Available from: <http://gbe.oxfordjournals.org/content/8/6/1906>
  53. Li L, Stoeckert CJ, Roos DS. OrthoMCL: Identification of Ortholog Groups for Eukaryotic Genomes. *Genome Research* [Internet]. 2003 Sep [cited 2017 Feb 8];13(9):2178–89. Available from: <http://genome.cshlp.org/content/13/9/2178>
  54. Waterhouse RM, Tegenfeldt F, Li J, Zdobnov EM, Kriventseva EV. OrthoDB:

- A hierarchical catalog of animal, fungal and bacterial orthologs. Nucleic Acids Research [Internet]. 2013 Jan [cited 2017 Feb 9];41(D1):D358–65. Available from: <https://academic.oup.com/nar/article/41/D1/D358/1060216/OrthoDB-a-hierarchical-catalog-of-animal-fungal>
55. Gao B, Gupta RS. Phylogenetic Framework and Molecular Signatures for the Main Clades of the Phylum Actinobacteria. Microbiology and Molecular Biology Reviews : MMBR [Internet]. 2012 Mar [cited 2017 Feb 8];76(1):66–112. Available from: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3294427/>
56. 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 [Internet]. 2014 [cited 2017 Feb 9];64(11):3821–32. Available from: <http://ijs.microbiologystresearch.org/content/journal/ijsem/10.1099/ijs.0.063966-0>
57. Zhou Z, Gu J, Li Y-Q, Wang Y. Genome plasticity and systems evolution in Streptomyces. BMC Bioinformatics [Internet]. 2012 [cited 2017 Feb 9];13(10):S8. Available from: <http://dx.doi.org/10.1186/1471-2105-13-S10-S8>
58. 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 Oct;25(10):1599–605.
59. 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 [Internet]. 2012 Aug [cited 2017 Feb 9];337(6098):1101–4. Available from: <http://science.sciencemag.org/content/337/6098/1101>
60. Copley SD. An Evolutionary Biochemist’s Perspective on Promiscuity. Trends in biochemical sciences [Internet]. 2015 Feb [cited 2017 Feb 8];40(2):72–8. Available from: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4836852/>
61. 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 [Internet]. 2013 Oct [cited 2017 Feb 9];502(7473):698–702. Available from: <http://www.nature.com/nature/journal/v502/n7473/full/nature12576.html>
62. Hughes AL. The Evolution of Functionally Novel Proteins after Gene Duplication. Proceedings of the Royal Society of London B: Biological Sciences [Internet]. 1994 May [cited 2017 Feb 8];256(1346):119–24. Available from: <http://rspb.royalsocietypublishing.org/content/256/1346/119>
63. Divergent Evolution of Enzymatic Function: Mechanistically Diverse Superfamilies and Functionally Distinct Suprafamilies. Annual Review of Biochemistry [Internet].

- 2001 [cited 2017 Feb 8];70(1):209–46. Available from: <http://dx.doi.org/10.1146/annurev.biochem.70.1.209>
64. 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* [Internet]. 2012 Feb [cited 2017 Feb 8];109(8):2966–71. Available from: <http://www.pnas.org/content/109/8/2966>
65. 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 Nov;69(5):512–26.
66. Merino E, Jensen RA, Yanofsky C. Evolution of bacterial trp operons and their regulation. *Current opinion in microbiology* [Internet]. 2008 Apr [cited 2017 Feb 10];11(2):78–86. Available from: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2387123/>
67. 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* [Internet]. 2016 May [cited 2017 Jan 31];473(9):1141–52. Available from: <http://www.biochemj.org/content/473/9/1141>
68. Lamble HJ, Heyer NI, Bull SD, Hough DW, Danson MJ. Metabolic Pathway Promiscuity in the Archaeon Sulfolobus solfataricus Revealed by Studies on Glucose Dehydrogenase and 2-Keto-3-deoxygluconate Aldolase. *Journal of Biological Chemistry* [Internet]. 2003 Sep [cited 2019 Jan 25];278(36):34066–72. Available from: <http://www.jbc.org/content/278/36/34066>
69. Weng J-K, Noel JP. The Remarkable Pliability and Promiscuity of Specialized Metabolism. *Cold Spring Harbor Symposia on Quantitative Biology* [Internet]. 2012 Jan [cited 2019 Jan 24];77:309–20. Available from: <http://symposium.cshlp.org/content/77/309>
70. Juárez-Vázquez AL, Edirisinghe JN, Verduzco-Castro EA, Michalska K, Wu C, Noda-García L, et al. Evolution of substrate specificity in a retained enzyme driven by gene loss. *eLife* [Internet]. [cited 2018 Jan 16];6. Available from: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5404923/>
71. Noda-García L. Estudio de la evolución molecular de la función enzimática susando como modelo una enzima con características ancestrales [PhD thesis]. [Irapuato, GTO]: Langebio, CINVESTAV; 2012.
72. 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* [Internet]. 2014 Jun [cited 2017 Feb 9];3:e03275.

- Available from: <https://elifesciences.org/content/3/e03275v2>
73. 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 [Internet]*. 2015 Jan [cited 2017 Feb 9];43(D1):D447–52. Available from: <https://academic.oup.com/nar/article/43/D1/D447/2435295/STRING-v10-protein-protein-interaction-networks>
  74. Segata N, Börnigen D, Morgan XC, Huttenhower C. PhyloPhlAn is a new method for improved phylogenetic and taxonomic placement of microbes. *Nature Communications [Internet]*. 2013 Aug [cited 2019 Jan 24];4:2304. Available from: <https://www.nature.com/articles/ncomms3304>
  75. 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 [Internet]*. 2015 [cited 2017 Jan 28];15(2):141–61. Available from: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4361730/>
  76. Koonin EV. The Turbulent Network Dynamics of Microbial Evolution and the Statistical Tree of Life. *Journal of Molecular Evolution [Internet]*. 2015 [cited 2017 Jan 28];80(5-6):244–50. Available from: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4472940/>
  77. Veen BE van der, Harris HM, O'Toole PW, Claesson MJ. Metaphor: Finding Bi-directional Best Hit homology relationships in (meta)genomic datasets. *Genomics [Internet]*. 2014 Dec [cited 2018 Jul 3];104(6, Part B):459–63. Available from: <http://www.sciencedirect.com/science/article/pii/S0888754314002092>
  78. Caetano-Anollés G, Yafremava LS, Gee H, Caetano-Anollés D, Kim HS, Mittenthal JE. The origin and evolution of modern metabolism. *The International Journal of Biochemistry & Cell Biology [Internet]*. 2009 Feb [cited 2018 Jul 3];41(2):285–97. Available from: <http://www.sciencedirect.com/science/article/pii/S1357272508003373>
  79. Schniete JK, Cruz-Morales P, Selem-Mojica N, Fernández-Martínez LT, Hunter IS, Barona-Gómez F, et al. Expanding Primary Metabolism Helps Generate the Metabolic Robustness To Facilitate Antibiotic Biosynthesis in Streptomyces. *mBio [Internet]*. 2018 Mar [cited 2018 Aug 10];9(1):e02283–17. Available from: <http://mbio.asm.org/content/9/1/e02283-17>
  80. Alanjary M, Kronmiller B, Adamek M, Blin K, Weber T, Huson D, et al. The Antibiotic Resistant Target Seeker (ARTS), an exploration engine for antibiotic cluster prioritization and novel drug target discovery. *Nucleic Acids Research [Internet]*. 2017 Jul [cited 2018 Jan 16];45(W1):W42–8. Available from: <https://academic.oup.com/nar/article/45/W1/W42/3787867>
  81. Martínez-Núñez MA, Rodríguez-Escamilla Z, Rodríguez-Vázquez K, Pérez-Rueda E. Tracing the Repertoire of Promiscuous Enzymes along the Metabolic Pathways in Archaeal Organisms. *Life [Internet]*. 2017 Jul [cited 2019 Jan 24];7(3). Available

- from: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5617955/>
82. Pearson H. Prehistoric proteins: Raising the dead. *Nature News* [Internet]. 2012 Mar [cited 2017 Feb 9];483(7390):390. Available from: <http://www.nature.com/news/prehistoric-proteins-raising-the-dead-1.10261>
  83. Treangen TJ, Rocha EPC. Horizontal Transfer, Not Duplication, Drives the Expansion of Protein Families in Prokaryotes. *PLOS Genetics* [Internet]. 2011 Jan [cited 2017 Feb 9];7(1):e1001284. Available from: <http://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1001284>
  84. 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* [Internet]. 1999 Mar [cited 2017 Feb 9];96(6):2896–901. Available from: <http://www.pnas.org/content/96/6/2896>
  85. 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* [Internet]. 2000 Sep [cited 2017 Feb 9];28(18):3442–4. Available from: [http://www.ncbi.nlm.nih.gov/pmc/articles/PMC110752/](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC110752/)
  86. Aziz RK, Bartels D, Best AA, DeJongh M, Disz T, Edwards RA, et al. The RAST Server: Rapid Annotations using Subsystems Technology. *BMC Genomics* [Internet]. 2008 Feb [cited 2017 Feb 7];9:75. Available from: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2265698/>
  87. 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* [Internet]. 2014 Jan [cited 2017 Feb 7];42(Database issue):D206–14. Available from: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3965101/>
  88. Harvey AL, Edrada-Ebel R, Quinn RJ. The re-emergence of natural products for drug discovery in the genomics era. *Nature Reviews Drug Discovery* [Internet]. 2015 Feb [cited 2016 Sep 16];14(2):111–29. Available from: <http://www.nature.com/nrd/journal/v14/n2/full/nrd4510.html>
  89. Petrenko R, Meller J. Molecular Dynamics. In: eLS [Internet]. John Wiley & Sons, Ltd; 2001 [cited 2017 Feb 8]. Available from: <http://onlinelibrary.wiley.com/doi/10.1002/9780470015902.a0003048.pub2/abstract>
  90. Molecular Modeling of Proteins Andreas Kukol Springer [Internet]. [cited 2017 Feb 8]. Available from: <http://www.springer.com/us/book/9781588298645>
  91. Sikosek T, Chan HS. Biophysics of protein evolution and evolutionary protein biophysics. *Journal of The Royal Society Interface* [Internet]. 2014 Nov [cited 2017 Feb 9];11(100):20140419. Available from: <http://rsif.royalsocietypublishing.org/>

- org/content/11/100/20140419
92. Zhou R. Replica Exchange Molecular Dynamics Method for Protein Folding Simulation. In: Bai Y, Nussinov R, editors. Protein Folding Protocols [Internet]. Humana Press; 2006 [cited 2017 Feb 9]. pp. 205–23. (Methods in Molecular Biology<sup>TM</sup>). Available from: <http://dx.doi.org/10.1385/1-59745-189-4%3A205>
  93. Bisswanger H. General Aspects of Enzyme Analysis. In: Practical Enzymology [Internet]. Wiley-VCH Verlag GmbH & Co. KGaA; 2011 [cited 2017 Feb 8]. pp. 5–91. Available from: <http://onlinelibrary.wiley.com/doi/10.1002/9783527659227.ch2/summary>
  94. Hommel U, Eberhard M, Kirschner K. Phosphoribosyl Anthranilate Isomerase Catalyzes a Reversible Amadori Reaction. Biochemistry [Internet]. 1995 Apr [cited 2017 Feb 8];34(16):5429–39. Available from: <http://dx.doi.org/10.1021/bi00016a014>
  95. Scheer M, Grote A, Chang A, Schomburg I, Munaretto C, Rother M, et al. BRENDA, the enzyme information system in 2011. Nucleic Acids Research [Internet]. 2011 Jan [cited 2017 Feb 9];39(suppl\_1):D670–6. Available from: [https://academic.oup.com/nar/article/39/suppl\\_1/D670/2506543/BRENDA-the-enzyme-information-system-in-2011](https://academic.oup.com/nar/article/39/suppl_1/D670/2506543/BRENDA-the-enzyme-information-system-in-2011)
  96. Van Der Spoel D, Lindahl E, Hess B, Groenhof G, Mark AE, Berendsen HJC. GROMACS: Fast, flexible, and free. Journal of Computational Chemistry [Internet]. 2005 Dec [cited 2017 Feb 9];26(16):1701–18. Available from: <http://onlinelibrary.wiley.com/doi/10.1002/jcc.20291/abstract>
  97. 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 [Internet]. 2014 Jun [cited 2017 Feb 9];111(23):8535–40. Available from: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4060685/>
  98. Osbourn A. Gene Clusters for Secondary Metabolic Pathways: An Emerging Theme in Plant Biology. Plant Physiology [Internet]. 2010 Oct [cited 2016 Sep 13];154(2):531–5. Available from: <http://www.plantphysiol.org/content/154/2/531>
  99. 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 [Internet]. 1999 Jul [cited 2017 Feb 12];9(7):608–28. Available from: <http://genome.cshlp.org/content/9/7/608>
  100. Benedict MN, Gonnerman MC, Metcalf WW, Price ND. Genome-Scale Metabolic Reconstruction and Hypothesis Testing in the Methanogenic Archaeon *Methanosarcina acetivorans* C2A. Journal of Bacteriology [Internet]. 2012 Feb [cited 2016 Sep 16];194(4):855–65. Available from:

- <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3272958/>
101. 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* [Internet]. 2016 Jul [cited 2016 Sep 16];10(7):1696–705. Available from: <http://www.nature.com/ismej/journal/v10/n7/full/ismej2015233a.html>
  102. Jeffries JG, Colastani RL, Elbadawi-Sidhu M, Kind T, Niehaus TD, Broadbelt LJ, et al. MINES: Open access databases of computationally predicted enzyme promiscuity products for untargeted metabolomics. *Journal of Cheminformatics* [Internet]. 2015 Dec [cited 2016 Sep 16];7(1). Available from: <http://www.jcheminf.com/content/7/1/44>
  103. Moustafa A, Loram JE, Hackett JD, Anderson DM, Plumley FG, Bhattacharya D. Origin of Saxitoxin Biosynthetic Genes in Cyanobacteria. *PLOS ONE* [Internet]. 2009 Jun [cited 2016 Sep 16];4(6):e5758. Available from: <http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0005758>
  104. Medema MH, Osbourn A. Computational genomic identification and functional reconstitution of plant natural product biosynthetic pathways. *Natural Product Reports*. 2016 Aug;33(8):951–62.
  105. Medema MH, Kottmann R, Yilmaz P, Cummings M, Biggins JB, Blin K, et al. Minimum Information about a Biosynthetic Gene cluster. *Nature Chemical Biology* [Internet]. 2015 Sep [cited 2017 Feb 12];11(9):625–31. Available from: <http://www.nature.com/nchembio/journal/v11/n9/full/nchembio.1890.html>
  106. Iqbal HA, Low-Beinart L, Obiajulu JU, Brady SF. Natural Product Discovery through Improved Functional Metagenomics in Streptomyces. *Journal of the American Chemical Society* [Internet]. 2016 Aug [cited 2017 Feb 12];138(30):9341–4. Available from: <http://dx.doi.org/10.1021/jacs.6b02921>
  107. 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* [Internet]. 2012 Aug [cited 2016 Sep 22];7(8). Available from: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3432047/>
  108. Charlesworth JC, Burns BP. Untapped Resources: Biotechnological Potential of Peptides and Secondary Metabolites in Archaea. *Archaea* [Internet]. 2015 Oct [cited 2016 Sep 27];2015:e282035. Available from: <http://www.hindawi.com/journals/archaea/2015/282035/abs/>
  109. Computational Pan-Genomics Consortium. Computational pan-genomics: Status, promises and challenges. *Briefings in Bioinformatics*. 2016 Oct;
  110. 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* [Internet]. 2015 Mar [cited 2016 Dec 7];6:6509. Available from: <http://www.nature.com/ncomms/2015/150305/ncomms7509/full/ncomms7509.html>
111. 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* [Internet]. 2013 Jul [cited 2016 Dec 7];499(7459):431–7. Available from: <http://www.nature.com/nature/journal/v499/n7459/full/nature12352.html?cookies=accepted>
112. 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* [Internet]. 2015 Mar [cited 2016 Dec 7];25(6):690–701. Available from: <http://www.sciencedirect.com/science/article/pii/S0960982215000160>
113. 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* [Internet]. 2015 May [cited 2016 Dec 7];521(7551):173–9. Available from: <http://www.nature.com/nature/journal/v521/n7551/full/nature14447.html>
114. Koonin EV. Archaeal ancestors of eukaryotes: Not so elusive any more. *BMC Biology* [Internet]. 2015 Oct [cited 2016 Dec 7];13. Available from: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4594999/>
115. Chaudhari NM, Gupta VK, Dutta C. BPGA- an ultra-fast pan-genome analysis pipeline. *Scientific Reports* [Internet]. 2016 Apr [cited 2016 Dec 7];6:24373. Available from: <http://www.nature.com/srep/2016/160413/srep24373/full/srep24373.html>
116. 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* [Internet]. 2006 Jan [cited 2016 Dec 7];103(2):425–30. Available from: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1324956/>
117. Narechania A, Baker RH, Sit R, Kolokotronics 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* [Internet]. 2012 Jan [cited 2017 Jan 23];4(1):30–43. Available from: <http://gbe.oxfordjournals.org/content/4/1/30>
118. Stamatakis A. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* [Internet]. 2014 May [cited 2016 Dec 22];30(9):1312–3. Available from: <http://www.ncbi.nlm.nih.gov/pmc/>

- articles/PMC3998144/
119. Powerful tree graphics with ggplot2 [Internet]. [cited 2017 Jan 28]. Available from: [http://joey711.github.io/phyloseq/plot\\_tree-examples.html](http://joey711.github.io/phyloseq/plot_tree-examples.html)
  120. Zacharia VM, Traxler MF. Exploring new horizons. *eLife* [Internet]. 2017 Jan [cited 2017 Jan 23];6:e23624. Available from: <https://elifesciences.org/content/6/e23624v1>
  121. Woese C. The universal ancestor. *Proceedings of the National Academy of Sciences* [Internet]. 1998 Jun [cited 2017 Feb 13];95(12):6854–9. Available from: <http://www.pnas.org/content/95/12/6854>
  122. Woese CR, Gupta R. Are archaebacteria merely derived “prokaryotes”? *Nature* [Internet]. 1981 Jan [cited 2017 Feb 13];289(5793):95–6. Available from: <http://www.nature.com/nature/journal/v289/n5793/abs/289095a0.html>
  123. 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 Jun;87(12):4576–9.
  124. 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* [Internet]. 1977 Nov [cited 2017 Jan 23];74(11):5088–90. Available from: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC432104/>
  125. Woese CR. There must be a prokaryote somewhere: Microbiology’s search for itself. *Microbiological Reviews* [Internet]. 1994 Mar [cited 2017 Jan 23];58(1):1–9. Available from: <http://mmbr.asm.org/content/58/1/1>
  126. Graham DE, Overbeek R, Olsen GJ, Woese CR. An archaeal genomic signature. *Proceedings of the National Academy of Sciences* [Internet]. 2000 Mar [cited 2017 Feb 13];97(7):3304–8. Available from: <http://www.pnas.org/content/97/7/3304>
  127. Howland JL. *The surprising archaea: Discovering another domain of life*. New York: Oxford University; 2000.
  128. Xu Y, Gogarten JP. *Computational Methods for Understanding Bacterial and Archaeal Genomes*. World Scientific; 2008.
  129. Garrett RA, Klenk H-P. *Archaea: Evolution, Physiology, and Molecular Biology*. John Wiley & Sons; 2008.
  130. Nishida H. Evolution of genome base composition and genome size in bacteria. *Frontiers in Microbiology* [Internet]. 2012 Dec [cited 2017 Jan 28];3. Available from: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3515811/>
  131. Coyle M, Hu J, Gartner Z. Mysteries in a Minimal Genome. *ACS Central Science* [Internet]. 2016 May [cited 2017 Feb 13];2(5):274–7. Available from:

- <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4882734/>
132. O'Meara B. CRAN Task View: Phylogenetics, Especially Comparative Methods. 2016 Dec [cited 2017 Jan 28]; Available from: <https://CRAN.R-project.org/view=Phylogenetics>
133. Larsson J, Nylander JA, Bergman B. Genome fluctuations in cyanobacteria reflect evolutionary, developmental and adaptive traits. BMC Evolutionary Biology [Internet]. 2011 [cited 2017 Feb 2];11:187. Available from: <http://dx.doi.org/10.1186/1471-2148-11-187>
134. Whitton BA. Ecology of Cyanobacteria II: Their Diversity in Space and Time. Springer Science & Business Media; 2012.
135. Cohen GN. The biosynthesis of histidine and its regulation. In: Microbial Biochemistry [Internet]. Springer Netherlands; 2004 [cited 2017 Jan 31]. pp. 225–30. Available from: [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)
136. 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 [Internet]. 2016 Jan [cited 2017 Feb 13];12(1):e1005836. Available from: <http://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1005836>
137. 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 [Internet]. 2004 [cited 2017 Feb 2];4:44. Available from: <http://dx.doi.org/10.1186/1471-2148-4-44>
138. Lapierre P, Gogarten JP. Estimating the size of the bacterial pan-genome. Trends in Genetics [Internet]. 2009 Mar [cited 2017 Feb 3];25(3):107–10. Available from: [http://www.cell.com/trends/genetics/abstract/S0168-9525\(09\)00005-5](http://www.cell.com/trends/genetics/abstract/S0168-9525(09)00005-5)
139. Větrovský T, Baldrian P. The Variability of the 16S rRNA Gene in Bacterial Genomes and Its Consequences for Bacterial Community Analyses. PLoS ONE [Internet]. 2013 Feb [cited 2017 Feb 8];8(2). Available from: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3583900/>
140. Labeda DP, Dunlap CA, Rong X, Huang Y, Doroghazi JR, Ju K-S, et al. Phylogenetic relationships in the family Streptomycetaceae using multi-locus sequence analysis. Antonie van Leeuwenhoek [Internet]. 2017 Apr [cited 2019 Mar 13];110(4):563–83. Available from: <https://doi.org/10.1007/s10482-016-0824-0>
141. Yasuhara-Bell J, Marrero G, Alvarez AM. Genes clvA, clvF and clvG are unique to *Clavibacter michiganensis* subsp. *michiganensis* and highly conserved. European Journal of Plant Pathology [Internet]. 2014 Dec [cited 2019 Mar 13];140(4):655–64.

- Available from: <https://doi.org/10.1007/s10658-014-0495-5>
142. 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* [Internet]. 2015 Feb [cited 2017 Feb 7];5:8365. Available from: <http://www.nature.com/srep/2015/150210/srep08365/full/srep08365.html>
  143. chesterismay. Updated R Markdown thesis template [Internet]. Chester's R blog. 2016 [cited 2017 Feb 7]. Available from: <https://chesterismay.wordpress.com/2016/09/01/updated-r-markdown-thesis-template/>
  144. Barona-Gómez F, Cruz-Morales P, Noda-García L. What can genome-scale metabolic network reconstructions do for prokaryotic systematics? *Antonie van Leeuwenhoek* [Internet]. 2012 Jan [cited 2017 Jan 24];101(1):35–43. Available from: <http://link.springer.com/article/10.1007/s10482-011-9655-1>
  145. 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. 265–83.
  146. Reed College. LaTeX your document [Internet]. 2007. Available from: <http://web.reed.edu/cis/help/LaTeX/index.html>
  147. Noble SG. Turning images into simple line-art [Undergraduate thesis]. Reed College; 2002.
  148. Angel E. Interactive computer graphics : A top-down approach with opengl. Boston, MA: Addison Wesley Longman; 2000.
  149. Angel E. Batch-file computer graphics : A bottom-up approach with quicktime. Boston, MA: Wesley Addison Longman; 2001.
  150. Angel E. Test second book by angel. Boston, MA: Wesley Addison Longman; 2001.