Enzymatic Promiscuity

A Thesis
Presented to
The Division of Mathematics and Natural Sciences
LANGEBIO-CINVESTAV

In Partial Fulfillment
of the Requirements for the Degree
Science Philosophy Doctor

Nelly Selem

Nov 2016

Approved for the Division
(Mathematics)

Francisco Barona Gomez

Acknowledgements

I want to thank a few people.

Preface

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

Table of Contents

List of Tables

List of Figures

Abstract

The preface pretty much says it all.

Second paragraph of abstract starts here.

Dedication

You can have a dedication here if you wish.

Background

Enzymes catalice chemical reactions transforming substrates into producs. During 20th century enzymes were percived as highly specific catalizer, nevertheless this percepcion changed with the discovery that they can . This abilty to catalice several chemical functions its known as enzyme promscuity. Escherichia coli contains at least 404 promiscuous enzymes. La relevancia de la promiscuidad radica tanto en su papel como mecanismo de evolucion de la funcion enzimatica, asi como en la necesidad de su deteccion para la correccion de modelos de flujo metabolico y la determinacion de efectos secundarios en drogas farmacologicas. A pesar de su frecuencia e importancia aun se esta en el proceso de entender las causas y las caracteristicas observables de la promiscuidad enzimatica.

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

Para mejorar nuestro entendimiento del fenomeno, ademas de la comparacion de secuencias es necesario integrar otros elementos de analisis. Se debe notar que es practicamente imposible decir que una enzima no es promiscua ya que para ello se deberian haber descartado todos los posibles sustratos. Sin embargo, para el estudio de cambios en promiscuidad se han detectado como elementos relevantes a los cambios en vecindad genomica, los cambios en flexibilidad durante la dinamica molecular, la perdida de genes centrales, y finalmente a las expansiones genicas dentro de un grupo taxonomico. Estos elementos tienen en comun que reflejan un cambio en alguna propiedad genomica o biofisica, de lo que se deriva que el buscar cambios en la promiscuidad de una enzima resulta mas factible que la busqueda intrinseca de promiscuidad, a lo que aspiran los metodos basados en comparaciones de secuencias.

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

0.1 Introduction

Para estudiar la promiscuidad es necesario contar con una definicion, algunos autores emplean el termino promiscuidad para describir actividades enzimaticas distintas a la funcion principal [1] otros lo ven como una actividad secundaria fortuita [2] que pudo aparecer de forma accidental o inducida artificialmente [3]. Otros mas, cuando una enzima puede operar sobre un amplio rango de sustratos, prefieren llamarla multiespecifica [1]. A la accion de realizar distintas funciones cataliticas, ya sea al catalizar varias reacciones quimicas o bien una misma reaccion en sustratos diferentes se le conoce como promiscuidad enzimatica [4]. Existen varios tipos de promiscuidad enzimatica.

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

Catalitica cuando la enzima utiliza diferentes mecanismos de reaccion y/o residuos cataliticos, e.g. la quimotripsina puede catalizar reacciones de amidasa y fosfotriesterasa en un mismo sitio activo. [4] Por condiciones del entorno, cuando la enzima cambia su conformacion dependiendo de las condiciones quimicas y fisicas presentes como pH, temperatura, solventes organicos y salinidad e.g. algunas lipasas pueden actuar como sintetizadoras de esteres en lugar de hidrolasas en presencia de solventes organicos [3].

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

asignan 4 digitos, los tres primeros corresponden a la reaccion y el ultimo al sustrato; el mayor numero de sustratos (4306 clases) que de reacciones quimicas (234 en el tercer nivel) sugiere que la mayor variacion evolutiva se da a nivel de sustrato y no de reaccion [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 comenzo 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.

TEX or LATEX
###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 probo 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 sobreexpresion 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.

0.1.1 Relacion del pangenoma con la promiscuidad enzimatica

El core genome de un grupo taxonomico es el conjunto de secuencias codificantes presentes en todos los organismos del grupo. En el Dominio Bacteria el core esta estimado entre 200 y 300 secuencias [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].

0.1.2 Expansion y contextos genomicos como herramienta de anotacion funcional

La diversidad enzimatica 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 [37]. Por este motivo detectar expansiones ya sea duplicaciones o transferencias horizontales [38], puede ser un buen punto de partida para determinar

divergencia funcional y promiscuidad. No todas las expansiones denotan cambio de funcion enzimatica, algunas pueden ser meros accidentes, sin embargo dado que la funcion de una enzima suele estar relacionada con sus vecinos [39], una expansion en una vecindad genomica diferente de la tradicional sera un referente de adquisicion de una nueva funcion y entonces un indicador de existencia previa de promiscuidad.

La funcion de una enzima es un concepto jerarquico, dependiente de la filogenia de un organismo [43]. Para sistematizar el estudio de contextos y vecindades genomicas se desarrollo Search Tool for the Retrieval of Interacting Genes/Proteins STRING [44], 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 [45] 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 [47]. 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 distintivo 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 [48], (Juarez Vazquez et al 2015).

0.1.3 Modelos bioinformaticos de promiscuidad

Con el fin de reducir la inversion en el proceso de experimentacion, se han implementado en los ultimos años algoritmos computacionales para predecir promiscuidad enzimatica [49]. 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 deteccion 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 caracteristicas 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 [48], lo que dificulta que considerar solo la secuencia conlleve a buenos predictores de promiscuidad. Cuando se obtiene una prediccion 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 catalizacion, 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).

0.1.4 Promiscuidad in vitro y promiscuidad in vivo

La ganancia de promiscuidad no solo puede entenderse como la capacidad de convertir mas sustratos [49], 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 Kcat,Kmson 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 [54] como los empleados en la deteccion de rutas no conservadas en la biosintesis de productos naturales [47]. En particular para este fin se ha utilizado espectrometria de masas MS/MS, [54] combinada con molecular networking para identificar productos similares [56]

0.1.5 El papel de la dinamica molecular en la promiscuidad

La estructura tridimensional de una proteina es obtenida mediante previa purificacion y cristalizacion. Aunque mucho se ha hablado de la relacion estructura funcion, al cristalizar se obtienen estados conformacionales homogeneos, que bien pueden no ser la unica conformacion que adopta la proteina en solucion. [58]. 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 [59]. Se postula que un aspecto de la dinamica molecular relevante para la diversificacion de especificidad por sustrato es el numero de conformeros [60]. 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 perdio 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 [48].

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 [61]. 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 β -lactamasas especialistas existentes [13].

0.1.6 Modelo biologico diversidad de Actinobacteria

Al escoger un conjunto acotado para investigar familias de enzimas promiscuas se debe recordar que la funcionalidad es jerarquica por lo que para mejorar la anotacion, es deseable reflejar el proceso evolutivo y restringirse a un grupo de organismos taxonomicamente relacionados [62]. 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 [63], en particular en Actinobacteria para identificar relaciones entre las familias del phylum, se obtuvieron arboles multilocus de entre 100 y 157 genomas [65]. Estos estudios sugieren como separar los genomas disponibles para hacer el calculo de grupos de ortologia. Finalmente, se han realizado estudios de plasticidad genomica en Streptomyces considerando 5 y 17 organismos de los 300

genomas disponibles en la actualidad [67] donde reportan 2,018 familias en el core genome y 32,574 en el pangenoma.

0.1.7 Modelo metabolico biosintesis de aminoacidos.

Al hacer el calculo vemos que Streptomyces, un genero del phylum Actinobacteria cuenta en su genoma con un promedio de 8316 secuencias codificantes segun la especie. Gran parte de estas secuencias pueden ser agrupadas en subsistemas metabolicos como metabolismo de carbohidratos o de lipidos; de estos subsistemas uno de los mas amplios es el metabolismo de aminoacidos con entre 429 y 910 secuencias segun el organismo. La sintesis de aminoacidos es un subsistema presente en todas las especies pero con suficientes variaciones que permiten hacer observaciones evolutivas. En un gran numero de Actinobacterias las rutas de histidina y triptofano de 7 y 11 pasos respectivamente convergen en una enzima bifuncional llamada PriA, que realiza tanto la funcion de HisA como la de TrpF [5]. La cantidad de familias en el subsistema de metabolismo de aminoacidos, su variabilidad, su conservacion entre distintos grupos taxonomicos y la existencia de estos ejemplos en Actinobacteria lo posicionan como un buen punto de partida para la busqueda de promiscuidad tanto de familias promiscuas como de miembros promiscuos de las mismas.

0.2 Antecedentes

En las cuatro decadas de estudio de la promiscuidad enzimatica, hemos aprendido que es un fenomeno distribuido en distintos subsistemas metabolicos [69] y que su existencia puede deberse tanto al desarrollo de nuevas funciones para fines adaptativos [17], como al rescate de una funcion perdida [27]. Por ello la dinamica de perdida y ganancia de genes asociada al contexto genomico en bacterias se relaciona con cambio en la funcion enzimatica [41]. Precisando, respecto a la ganancia de genes, se postula que la bifuncionalidad precede la duplicacion [37]. Lo que implica que dada una duplicacion muy posiblemente previamente la promiscuidad estuvo presente [71].

Se han desarrollado tecnicas bioquimicas y metabolicas de medicion [12], asi como algoritmos computacionales de prediccion de promiscuidad [49]. Un aspecto a mejorar dentro del modelado es la restriccion del conjunto de estudio a un grupo taxonomico tan reducido que exista congruencia en las familias de ortologia y a la vez tan amplio que permita observar efectos evolutivos; el phylum Actinobacteria ha probado tener ejemplos de promiscuidad. Si bien la secuencia no ha sido suficiente para la correcta prediccion de promiscuidad [42], es posible que dentro de las tecnicas computacionales la flexibilidad durante la dinamica molecular este correlacionada con la promiscuidad de los miembros de una familia [58].

0.2.1 Modelo biologico

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`), uno de ellos correspondiente a Streptomycetaceae, la familia con la mayor cantidad de genomas disponibles; los otros dos grupos siguieron la taxonomia propuesta por Gao & Gupta en 2012. En el grupo de 290 genomas de Streptomycetaceae 2,126,832 ORFS fueron clasificados en 288,390 familias; de las 919,292 ORF del grupo I de Actinobacteria resultaron 269,406 familias. Las relaciones taxonomicas fueron corroboradas con algoritmos propios basados en best bidirectional hits (BBH).

0.2.2 Subsistemas metabolicos

Los operones his y trp de histidina [73] y triptofano [74] 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 [75]. 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 [48] y en la familia subtrpF se conserva solo a la funcion TrpF debido a la perdida del operon his [Juarez vazquez et al 2015 in prep]. Existen al menos 43 familias de Actinobacteria sin explorar respecto a la funcionalidad de PriA.

0.2.3 Contexto y vecindades genomicas

En 2012 fueron analizados 102 genomas de 29 familias de Actinobacteria [76]. sugiriendo que al menos en Corynebacteria el contexto y la vecindad genomica incidian en la sub-funcionalizacion de PriA en subHisA [48]. 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]

0.2.4 Metodos bioinformaticos

Al evaluar PROMISE [49] en un set de datos de la familia HisA/PriA [52] obtuve que su mejor desempeño es con huella molecular de tamaño 6, donde clasifica correctamente casi todas las no promiscuas, (HisA) pero no sucede lo mismo con la familia PriA donde tiene exito en 16 de 45 casos. Al aplicar el mismo tamaño de huella a 9 miembros promiscuos de la familia IlvC no consigue predecir correctamente ninguno de ellos. Por lo menos para estas familias el conjunto de entrenamiento o los descriptores no son suficientes para la anotacion de promiscuidad.

0.2.5 Evomining

Evomining es una plataforma bioinformatica pensada para la identificacion de productos naturales que tiene entre sus exitos la identificacion de la biosintesis de arsenolipidos [62]. La busqueda de productos naturales cuenta entre sus premisas que estos se producen en vecindades genomicas llamadas clusters y que ademas clusters cercanos (ya sea en contenido genico o en la secuencia de sus componentes), exploran variaciones metabolicas, es decir sus enzimas catalizan reacciones sobre sustratos parecidos aunque no identicos [62]. La base de evomining es que las enzimas de metabolismo secundario son expansiones distantes de enzimas de rutas centrales, lo que da idea de la quimica que realizan dichas expansiones dejando por identificar el sustrato sobre el que trabajan. La primera version de evomining cuenta con 200 genomas de Actinobacteria, una base de datos de secuencias de enzimas de productos naturales y otra base de datos de secuencias de enzimas de rutas centrales curada a mano. Evomining esta ligada con el problema de la promiscuidad porque en estas familias expandidas ya sea por duplicacion o por transferencia horizontal, las expansiones pueden retener la funcion quimica de las rutas centrales y viceversa, la funcion quimica expandida suele estar presente antes de la duplicacion.

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

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

0.2.6 Caracterizacion in vivo

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

0.2.7 Caracterizacion bioquimica in vitro.

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

0.2.8 Modelado de dinamica molecular

La dinamica es un metodo que permite hacer simulaciones de particulas que sirve para obtener informacion de propiedades macroscopicas de un conjunto de atomos [77]. 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 [79], 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[55]. 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 [52], 30 estructuras y dinamicas de miembros de Actinobacteriaceae y Bifidobacteriaceae (Vazquez-Juarez et al in prep.) y finalmente una estructura de subHisA de Corynebacterium diphteriae. 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 [76] se utilizo el metodo de dinamica molecular y se comparo el numero de conformeros entre miembros de subHisA y PriA, resultando mayor el de PriA como corresponde a una enzima promiscua. El estudio sobre la relacion dinamica-flexibilidad de β-lactamasas utiliza replica exchange, una variacion de dinamica molecular que corre replicas en paralelo a distintas temperaturas [80]. Una desventaja de este metodo es que por el costo computacional de las replicas agregar explicitamente otras moleculas a la simulacion como el solvente no es posible en tiempo razonable. Una vez generadas las dinamicas moleculares se procedera a calcular tanto el numero de conformeros como el indice de flexibilidad dsi [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 de familias de enzimas promiscuas, sin embargo hasta ahora no se han reportado trabajos multidisciplinarios que involucren a todas las partes involucradas (Fig 6)

0.3 Objetivo General

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

0.4 Objetivos particulares

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

0.5 Estrategias

0.5.1 La promiscuidad en familias enzimaticas.

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

Obtener informacion genomica del phylum Actinobacteria.

Colectar genomas de Actinobacteria de NCBI y de colecciones privadas.
###Anotar consistentemente las secuencias codificantes de estos genomas.
Utilizar un anotador automatizado y desarrollar los scripts necesarios para anotar los genomas.
####Establecer las relaciones filogeneticas de los genomas colectados.
Mediante el uso del core genome construir un arbol filogenomico que permita establecer un marco sobre el cual hablar de cambio y que facilite reclasificar los genomas mal nombrados.
#####Identificar cambios en la vecindad genomica en familias selectas de enzimas de metabolismo central. Clasificar sistematicamente las secuencias de familias codificantes segun su similitud en familias enzimaticas.
Desarrollar las herramientas bioinformaticas necesarias para separar clusters de vecindades genomicas.
####Sistematizar Evomining para convertirla una plataforma descargable y utilizable en cualquier set de datos bacterianos relacionados taxonomicamente proporcionados por el usuario.
Ampliar el contenido de Evomining al integrar los genomas colectados de Actinobacteria. Sistematizar la base de datos de metabolismo central.
Desarrollar la visualizacion e integrar la clasificacion de vecindades genomicas como una herramienta adicional en la busqueda de promiscuidad.

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

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

Seleccionar miembros homologos de la familia de enzimas.

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

cuenta con cinetica enzimatica de PriA, tres de ellos con estructura cristalografica.
Medir cineticas enzimaticas, contexto genomico, vecindad genomica, flexibilidad y numero de conformeros.
Determinar la pertenencia a uno de cuatro posibles contextos genomicos respecto al gen trpF. Estudiar la existencia de distintas vecindades genomicas. Determinar la cinetica enzimatica de 9 enzimas mas buscando variabilidad en contexto genomico (sugeridas en la tabla 4). Obtener mediante una colaboracion 37 modelos estructurales por homologia y modelar dinamica molecular.

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

Determinar posibles correlaciones entre los datos producidos.

Numero de conformeros e indice de promiscuidad.
indice de flexibilidad y numero de conformeros.
Numero de conformeros y contexto genomico.
indice de flexibilidad y contexto genomico.
Contexto genomico e indice de promiscuidad I.
Analizar las vecindades genomicas e indice de promiscuidad I.

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

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

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

Dado un organismo modelo sustituir su homologo nativo de priA por una variante no nativa ya sea de priA o trpF de Actinobacterias selectas de las que se sospecha cambio en promiscuidad.
####Separar posibles productos y minimizar los falsos positivos debidos a perturbaciones metabolicas no relacionadas a PriA.
Separar los metabolitos mediante cromatografia dirigida por tamaño. Obtener un espectro de masas antes y despues de la sustitucion de la variante y sobre las diferencias en el espectro realizar espectrometria de masas en tandem (MS/MS) es decir refragmentar y analizar que los fragmentos contengan partes parecidas a los sustratos conocidos.

0.6 Metodologia

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

0.6.1 La promiscuidad en familias enzimaticas.

Actinobacteria genomic

Para obtener informacion genomica del phylum Actinobacteria mediante la coleccion 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 deteccion de una familia marcadora de Clavibacter michiganensis (2014 Rodriguez-Orduña in prep).

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

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

Como se menciono en los antecedentes, se han separado 888 genomas de Actinobacteria en 3 grupos taxonomicos utilizando para la anotacion la tecnologia de subsistemas de RAST. Para la separacion en familias iso funcionales (ortologos, paralogos y expansiones) n se utilizo RAST, especificamente el script What Changed (WC) que asigna un numero a cada familia, esta herramienta esta basada en k-mers, su codigo esta disponible en github: (`https://github.com/kbase/kbseed/blob/master/service-scripts/svr_CS.pl`). 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 definicio 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 Aes n-similar a otra vecindad B si C={aA | b B ,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 esn-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.

```
#$$   M\eq\max_{ni} i\eq{1,2,..,k}          m\eq\max_ni i\eq {1,2,...,k}  ni\ne{M
```

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

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

La evaluacion de Cl esta dada por una combinacion lineal de k,m y M
S(Cl)=f(k,m,M)=c1k+c2m+c3M

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

Organizar y presentar los datos en una plataforma.

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

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

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

0.6.3 Promiscuidad in vitro

Datos cineticos:

En todos los ensayos enzimaticos se busca medir una señal que permita una distincion clara entre sustrato y producto [81]. 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 [82]. 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 [75].

Finalmente se recopilaran datos cineticos de PriA tanto privados como los publicos reportados a la fecha en la BRaunschweig ENzyme DAtabase BRENDA [83]. 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=1ln Ni=1Npi(pi) donde pi=Kicat Kim / i=1NKicatKim

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

0.6.4 Promiscuidad in vivo

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

se pasara por conjugacion a S. coelicolor donde se espera que realice una doble recombinacion. El paso por E. coli es llevado a cabo porque Streptomyces no se puede transformar por electroporacion. Se seleccionaran las cepas de Streptomyces resistentes al antibiotico como prueba de que ya no poseen su priA nativa. Posteriormente, mediante un procedimiento analogo se 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 moleculas. Basicamente estos metodos se basan en diferencias en el tamaño, intercambio de iones y afinidad. [55] Posteriormente se combinan con espectrometria de masas que es una tecnica que mide el radio masa-carga de las particulas fragmentadas en iones. [55]. Los datos obtenidos de espectrometria de masas se procesaran utilizando redes moleculares, que consiste en agrupar los productos segun la similitud de sus partes. Plan: 3 replicas tecnicas, 2 replicas biologicas de 5 cepas.

0.6.5 Consideraciones

Falsos negativos respecto a promiscuidad estan muy extendidos en la literatura y en las bases de datos, en parte porque la mayoria de las funciones son asignadas por similitud de secuencia y dado un falso negativo el error se propaga en secuencias similares. Por otro lado es muy dificil demostrar un verdadero negativo a menos que se prueben todas las posibilidades de sustrato para la enzima. Sin embargo el espacio de sustratos puede acotarse gracias a tecnicas como el docking que esta intimamente relacionado con la dinamica molecular [55]. Limitar el espacio de sustratos puede retroalimentarse con el estudio de la promiscuidad in vivo y viceversa.

Con los metodos propuestos en este trabajo solo se podra detectar perdida o ganancia de promiscuidad entre enzimas de organismos respecto a otros miembros dentro un grupo taxonomico, no asi el estado de promiscuidad intrinseco a la enzima. Si dada una enzima no se detectan variaciones en contexto, vecindad genomica o flexibilidad dentro de un grupo taxonomico cercano, entonces no podemos decir en principio nada acerca de la promiscuidad de la variante, posiblemente es promiscua pero al mantenerse constante en todos los parametros descritos, con estos metodos no se puede sugerir promiscuidad. Es posible que al mirar en un grupo taxonomico mas amplio se detecte una neofuncionalizacion de la familia aunque tambien es posible que exista una variable z como la flexibilidad de sustrato [20] que no se este considerando y que explique o sea el mejor indicador para esta familia de promiscuidad enzimatica.

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

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

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

Gene cluster plants[86]
Archaeal core [87]
Natural products genomic era[88]
Methanosarcina reconstruction [89]
Archaea phylum[90]
Prediction for possible products of promiscuous enzymes[**???**]
Saxitoxin [91]
Plants clusters [92]
MiBIG [93]
Metagenomics on Streptomyces [94]
Sulfolobus reconsruction [95]
Archaeal Natural products[96]
Computational Pangenomics [97]
Cuántos genes "obtenidos por EvoMining" son core/ cloud/stand alone
Qué porcentaje de genes únicos recupera EvoMining
Eucarya paralogs reshape gene clusters [98]
Microbial dark mater [99]
Archaea anaerobica carbon [100] Archaea Eucarya gap loki[101]
Archaea and eucarya[102]
BPGA [103] genes esenciales bacteria minima[104]
Radical [105]
RaxML large philogenies [106]
R phylogenies [107]
Streptomyces exploradores [108]
LUCA [109] Luchando por el reconocimiento de Archaea[110],[111] The primary kingdoms [112]
Predicccion aRchaeas [113]
RASt archaea [114] Book Archaea [115]
Computational methods for bacterial and archaeal genomes [116]
Archaeas boook [117]
Bacterial /archaeal genome [118]
Bacteria Archaea genome [119]
Tree of life and HGC [120]
Genomas retrospectiva 20 años [121]
GC content plasmido genoma [122] Genoma minimo[123]
Phylogeny R [124]
Cyanobacteria fluctuacion genomica y adaptacion [125]
Ecology of cyanobacterua [126]
Histdine biosynthesis[127]
PriA reconstruction [128]

Escala temporl bacterias [129]
Pangenome size [130]
variabilidad del 16s [131]

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

Chapter 1

EvoMining

1.1 Introduction

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

1.2 Gen families expansions on genomes

1.2.1 Pangenomes

Expansions are located on pangenome, Tools to analyse pangenome BPgA

1.3 EvoMining

EvoMining looks expansions on prokariotic pangenome.
Biological idea.

EvoMining was available as a consult website with 230 members of the Actinobacteria phylum as genomic data base, 226 unclassified nBGCs, and not interchangable central database 339 queries for nine pathways, including amino acid biosynthesis, glycolysis, pentose phosphate pathway, and tricarboxylic acids cycle. [62] EvoMining was proved on Actinobacteria Arseno-lipids

1.4 Pangenome

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

HGT doubles mutation rate on prokarites.
Maybe HGT is an selected feature, if is the case, so could be np production.
Some archaeas has open pangenome. [34]

HGT doubles mutation rate on prokarites. [120] Maybe HGT is an selected feature, if is the case, so could be np production.
Some archaeas has open pangenome. [34] Shell trees converge to core trees [105]

1.5 EvoMining Implementation

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

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

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

Dockerhub is an apps project hosting.

Dockerhub nselem

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

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

a DockerHub automated build repository linked to github EvoMining repository so that code is always actualized.

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

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

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

Other genomic apps were containerized to docker images during this work.
- *myRAST* docker- `https://github.com/nselem/myrast`
RAST is a bacterial and Archaeal genome annotator [45] This app allows myRAST functionality to upload
It allows EvoMining genome database annotation.
-*Orthocores* docker-`https://github.com/nselem/orthocore`
Helps to obtain genomic core paralog free and construct genomic trees
-*CORASON* docker-`https://github.com/nselem/EvoDivMet/wiki`
-PseudoCore github- <>
Genomic Core with a reference genome has the advantage of more genomes, but it is not paralog free

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

EvoMining Dockerization was chosen to avoid future compatibilty problems, for example dependencies unavailabilty, or incompatibility between future versions of its software components. As much as reproducible research was a concerned while developing EvoMining app, reproducibilty is also important on data analysis, for that reason this document was writen using R-markdown and latex template from Reed College [133]. While R-markdown allows to write and run R code and interpolate text paragraph to explain scripts and analysis.

1.6 EvoMining Databases

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

1. Genomes DB
2. Natural Products DB

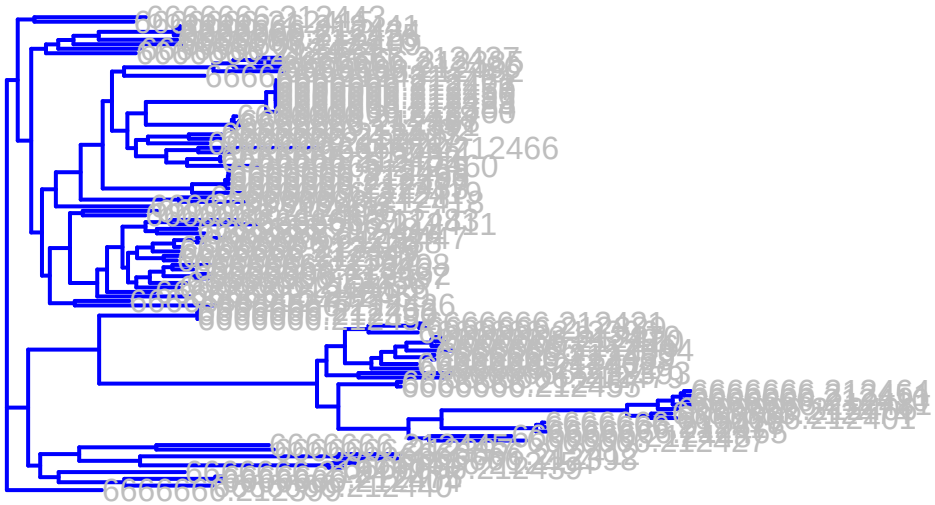
3. Central Pathways DB

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

Genome DB

RAST annotation of genomes was done.

Phylogeny



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

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

General Trees
Actinobacteria Tree, ArchaeaTree, CyanobacteriaTree.

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

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

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

Central DB

We chose central pathways from [134]
* BBH Best Bidirectional Hits with studied enzymes from Central Actinobacterial pathways were selected.

- By abundance
- By expansions on genomes

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

1.7 Data Bases

1.7.1 Central pathways

Central database were chosen by BBH from

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

Table 1.1: BBH_Organisms

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

1.7.2 Genome Dynamics

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

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

Selected organims were

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

Table 1.2: WC_Organisms

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

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

Those families expanded on at least two groups

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

Those Families expanded on Archaea and not expanded on Actino

```
comm -23 f3Archaeas f3Actinos >ArchaeasNoActinos
```

Those Families expanded on Actino and not on Archaea

```
comm -13 f3Archaeas f3Actinos >ActinosNoArchaea
```

Those families expanded on Streptomyces but not in ActinoBacteria

```
comm -13 f343Strepto f3Actinos >ActinosNoStrepto
```

Those Families expanded on Actinobacteria and not in Streptomyces

```
comm -23 f343Strepto f3Actinos >StreptoNoActinos
```

Those Families expanded on Cyano and not in Actino

```
comm -23 f3Cyanos f3Actinos >CyanosNoActinos
```

Natural Products DB

Natural products was improved from previous version

1.7.3 AntisMASH optional DB

AntiSMASH is [135]
Archaeas Results Archaea is a kingdom of recent discovery were not many natural products has been known. On Actinobacteria, evoMining has proved its value to find new kinds of natural products. The clue to this discovery was that Actinobacteria has genomic expanssions. Now Archaea has genomic expansions, even more has central pathways genomic expansions. Are this expansions derived from a genomic duplication?
Has Archaea natural products detected by antismash, and if not, where are this NP's or may Archaea doesn't have NP's.

applying EvoMining to Archaea

1.7.4 Otras estrategias para los clusters Argon context Idea

1.8 Argonne

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

cs close strain
wc whats chain

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

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

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

fig|6666666.112839.peg.26

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

cluster_objects < elements_to_cluster > ClusteFile

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

write a file similiar but with the family number

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

compare each peg on this file from the same family

Write the conextions file
peg1 peg2
peg1 peg3
peg2 peg3

cluster this file and score the cluster

Define

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

as a “family”

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

members span the sets

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

that make up the

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

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

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

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

cluster these context using the "similar contexts" relation

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

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

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

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

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

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

history

Para hacer un nuevo set de datos

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

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

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

```
summary(cars)
```

```
     speed           dist
 Min.   : 4.0   Min.   :  2.00
```

```
 1st Qu.:12.0   1st Qu.: 26.00
 Median :15.0   Median : 36.00
 Mean   :15.4   Mean   : 42.98
 3rd Qu.:19.0   3rd Qu.: 56.00
 Max.   :25.0   Max.   :120.00
```

1.8.1 Inline code

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

> The `cos` of 2π is 1.

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

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

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

> The standard deviation is less than 6.

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

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

1.9 Recomendaciones de Luis

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

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

2. dist tree
   quarter descomposition
   peter gogarten fendou Mao

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

Para obtener
Pablo Vinuesa: Get Homologues

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

Servidor Viernes a las 12:00

1.10 CORASON: Other genome Mining tools context-based

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

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

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

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

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

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

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

1.12 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*

```
{r chapter2, child = 'chap2.Rmd'}
```

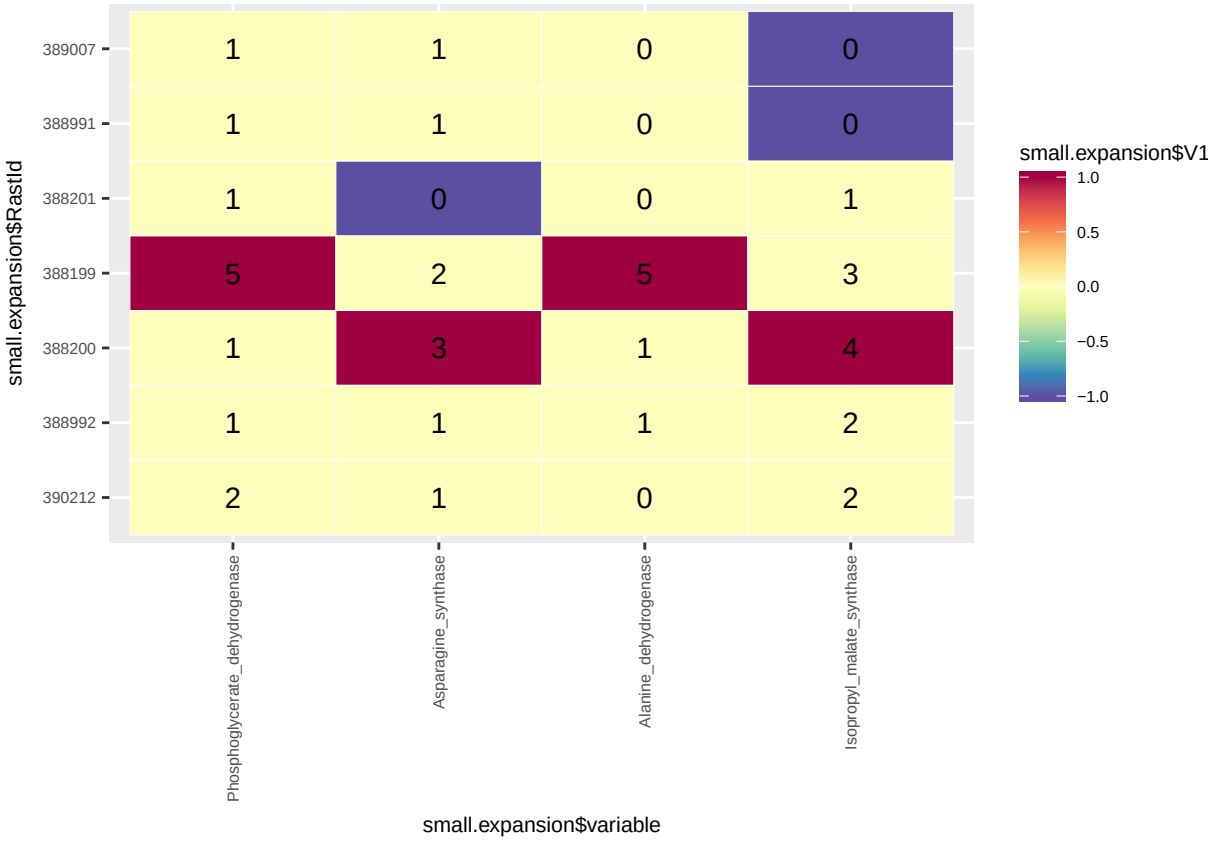
```
ArchaeasHeatPlot <- read.table("chapter3/ArchaeasHeatPlot", header=TRUE, sep="\t")
ArchaeasTaxa <- read.table("chapter3/ArchaeasTaxa", header=TRUE, sep="\t")

## Adding order variable
ArchaeasHeatPlot$order<-c(1:nrow(ArchaeasHeatPlot))
```

```
#sorting RastId it accordig to order variable
ArchaeasHeatPlot$RastId <- with(ArchaeasHeatPlot,reorder(ArchaeasHeatPlot$RastId,

# Merging heatplot and taxonomy table into one table
HP_Archaeas_Taxa<-merge(ArchaeasHeatPlot,ArchaeasTaxa,by.x = "RastId",by.y = "Rast
HP_Archaeas_Taxa.m<- melt(HP_Archaeas_Taxa)
```

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

```
HP_Archaeas_Taxa.m<- ddply(HP_Archaeas_Taxa.m, .(variable), transform,rescale=scal

small<-HP_Archaeas_Taxa.m[which(HP_Archaeas_Taxa.m$RastId=='389007' | HP_Archaeas_

#Este Escogiendo que variables quiero
small<-small[small$variable %in% c("Phosphoglycerate_dehydrogenase", "Isopropyl_ma
## use http://colorbrewer2.org/ to find optimal divergent color palette (or set ow

small.m<- ddply(small, .(variable),summarize, mean = round(mean(value), 2),sd = ro
rownames(small.m)<-small.m$variable
```

```
color_exp<-function(x){
  # Pendiente pasarle un dataframe en lugar de tener fijo small.m como variable
  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<-small.m$expansion [which(small.m$variable==met_family)]
  reduction<-small.m$reduction [which(small.m$variable==met_family)]
  if (local_value>=expansion){result= 1}
  else if (local_value<=reduction){result= -1}
  return (result)
}

small.expansion<-adply(small,1,color_exp)

#small.expansion$V1
#color_palette <- colorRampPalette(c("#3794bf", "#FFFFFF","#df8640"))

ggplot(small.expansion, aes(small.expansion$variable, small.expansion$RastId,label
```




```
############################################################
## Trying to sort the heatplot
## Reading heatplot table and taxa information and saving it on data.frame data structur
ArchaeasHeatPlot <- read.table("chapter3/ArchaeasHeatPlot", header=TRUE, sep="\t")
ArchaeasTaxa <- read.table("chapter3/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, Archae
# Merging heatplot and taxonomy table into one table
HP_Archaeas_Taxa<-merge(ArchaeasHeatPlot,ArchaeasTaxa,by.x = "RastId",by.y = "RastId")
# Melting information leting as variables just enzymatic families and copy number
HP_Archaeas_Taxa.m<- melt(HP_Archaeas_Taxa)
```

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

```
# Cleaning data
HP_Archaeas_Taxa.m<- ddply(HP_Archaeas_Taxa.m, .(variable), transform,value)  ##
HP_Archaeas_Taxa.m<- ddply(HP_Archaeas_Taxa.m, .(variable), transform,rescale=scale(valu

HP_Archeas.calcs<- ddply(HP_Archaeas_Taxa.m, .(variable),summarize, mean = round(mean(va
```

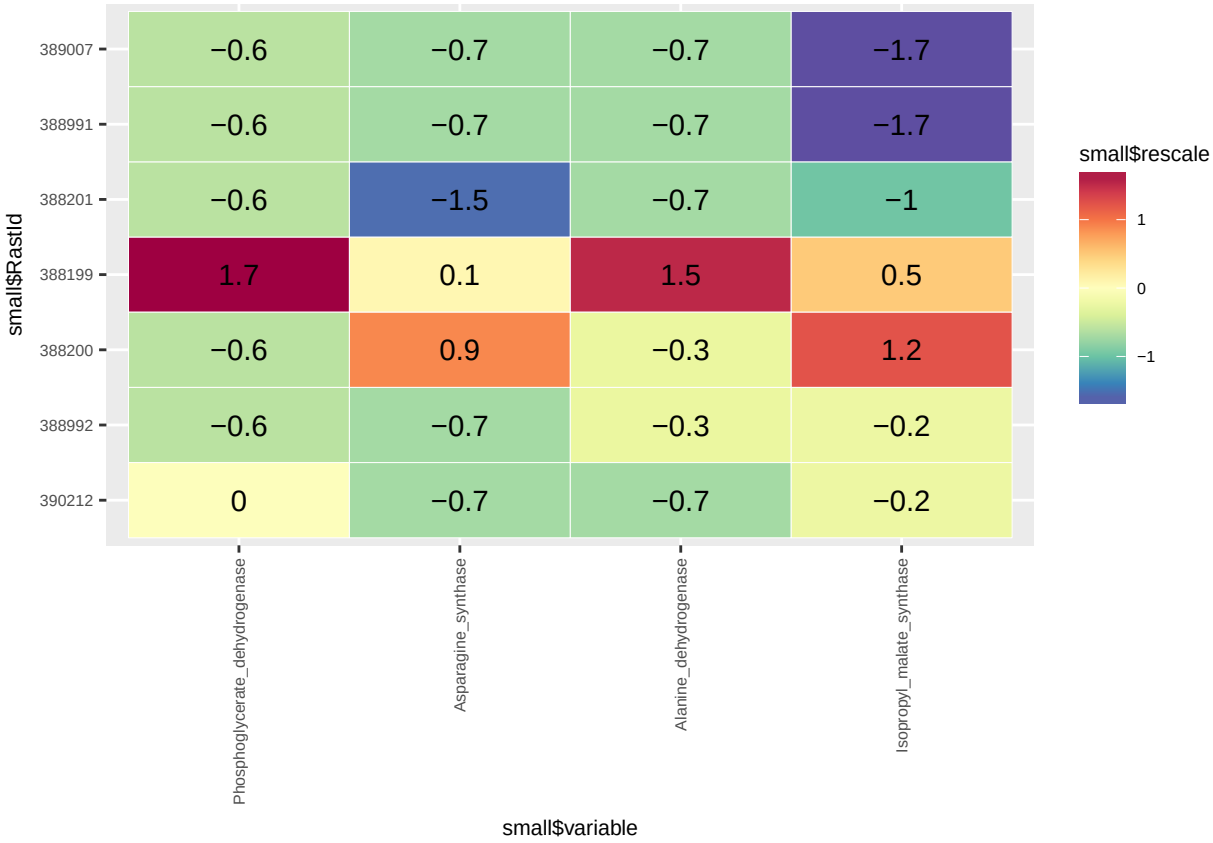
```
rownames(HP_Archeas.calcs)<-HP_Archeas.calcs$variable
############################################################3
color_exp<-function(x){
  # Pendiente pasarle un dataframe en lugar de tener fijo small.m como variable
  result=0
  expansion<-NULL
  reduction<-NULL
  local_value=x[1,"value"]
  met_family<-x[1,"variable"]
  Rast=x[1,1]
  #print ("rast",Rast)
  # print(paste(x,"family",met_family,"value",local_value,"Rast",Rast))
  expansion<-HP_Archeas.calcs$expansion [which(small.m$variable==met_family)]
  reduction<-HP_Archeas.calcs$reduction [which(small.m$variable==met_family)]
  if (local_value>=expansion){result= 1}
  else if (local_value<=reduction){result= -1}
  return (result)
}
small[ !small$variable %in% c("Contigs", "Size","TOTAL"), ]
```

	RastId	Name
1	388199	Haloferax sp ATCC BAA-644ATCC BAA-644 AOLF01
2	388200	Candidatus Methanoperedens nitroreducensANME-2d JMIY01
3	388201	Marine group II euryarchaeote REDSEA-S40_B11N13 LURX01
551	388991	Uncultured Acidilobus sp MG AYMA01
552	388992	Candidate divison MSBL1 archaeon SCGC-AAA259O05 LHXV01
567	389007	Uncultured Acidilobus sp OSP8 AYMC01
690	390212	Archaeoglobus fulgidus DSM 8774 DSM 8774 CP006577.1
14017	388199	Haloferax sp ATCC BAA-644ATCC BAA-644 AOLF01
14018	388200	Candidatus Methanoperedens nitroreducensANME-2d JMIY01
14019	388201	Marine group II euryarchaeote REDSEA-S40_B11N13 LURX01
14567	388991	Uncultured Acidilobus sp MG AYMA01
14568	388992	Candidate divison MSBL1 archaeon SCGC-AAA259O05 LHXV01
14583	389007	Uncultured Acidilobus sp OSP8 AYMC01
14706	390212	Archaeoglobus fulgidus DSM 8774 DSM 8774 CP006577.1
63073	388199	Haloferax sp ATCC BAA-644ATCC BAA-644 AOLF01
63074	388200	Candidatus Methanoperedens nitroreducensANME-2d JMIY01
63075	388201	Marine group II euryarchaeote REDSEA-S40_B11N13 LURX01
63623	388991	Uncultured Acidilobus sp MG AYMA01
63624	388992	Candidate divison MSBL1 archaeon SCGC-AAA259O05 LHXV01
63639	389007	Uncultured Acidilobus sp OSP8 AYMC01
63762	390212	Archaeoglobus fulgidus DSM 8774 DSM 8774 CP006577.1
68329	388199	Haloferax sp ATCC BAA-644ATCC BAA-644 AOLF01
68330	388200	Candidatus Methanoperedens nitroreducensANME-2d JMIY01

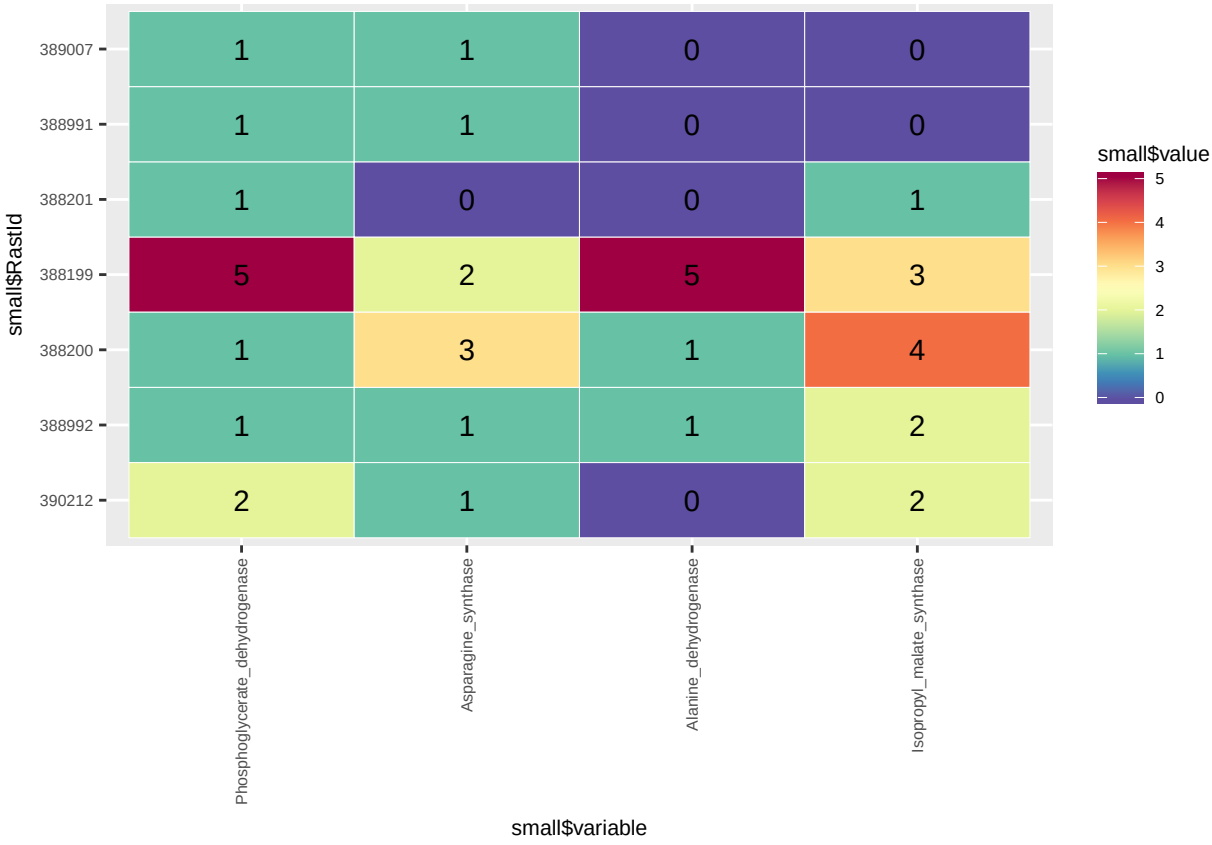
```
68331 388201 Marine group II euryarchaeote REDSEA-S40_B11N13 LURX01
68879 388991                     Uncultured Acidilobus sp MG AYMA01
68880 388992 Candidate divison MSBL1 archaeon SCGC-AAA259O05 LHXV01
68895 389007                   Uncultured Acidilobus sp OSP8 AYMC01
69018 390212    Archaeoglobus fulgidus DSM 8774 DSM 8774 CP006577.1
        SuperPhylum        Phylum           Class             Order
1     Euryarchaeota Euryarchaeota    Halobacteria     Haloferacales
2     Euryarchaeota Euryarchaeota Methanomicrobia Methanosarcinales
3     Euryarchaeota Euryarchaeota    unclassified      unclassified
551      TACK group Crenarchaeota    Thermoprotei      Acidilobales
552   Euryarchaeota Euryarchaeota    unclassified      unclassified
567      TACK group Crenarchaeota    Thermoprotei      Acidilobales
690   Euryarchaeota Euryarchaeota    Archaeoglobi   Archaeoglobales
14017 Euryarchaeota Euryarchaeota    Halobacteria     Haloferacales
14018 Euryarchaeota Euryarchaeota Methanomicrobia Methanosarcinales
14019 Euryarchaeota Euryarchaeota    unclassified      unclassified
14567    TACK group Crenarchaeota    Thermoprotei      Acidilobales
14568 Euryarchaeota Euryarchaeota    unclassified      unclassified
14583    TACK group Crenarchaeota    Thermoprotei      Acidilobales
14706 Euryarchaeota Euryarchaeota    Archaeoglobi   Archaeoglobales
63073 Euryarchaeota Euryarchaeota    Halobacteria     Haloferacales
63074 Euryarchaeota Euryarchaeota Methanomicrobia Methanosarcinales
63075 Euryarchaeota Euryarchaeota    unclassified      unclassified
63623    TACK group Crenarchaeota    Thermoprotei      Acidilobales
63624 Euryarchaeota Euryarchaeota    unclassified      unclassified
63639    TACK group Crenarchaeota    Thermoprotei      Acidilobales
63762 Euryarchaeota Euryarchaeota    Archaeoglobi   Archaeoglobales
68329 Euryarchaeota Euryarchaeota    Halobacteria     Haloferacales
68330 Euryarchaeota Euryarchaeota Methanomicrobia Methanosarcinales
68331 Euryarchaeota Euryarchaeota    unclassified      unclassified
68879    TACK group Crenarchaeota    Thermoprotei      Acidilobales
68880 Euryarchaeota Euryarchaeota    unclassified      unclassified
68895    TACK group Crenarchaeota    Thermoprotei      Acidilobales
69018 Euryarchaeota Euryarchaeota    Archaeoglobi   Archaeoglobales
                 Family                         variable value     rescale
1          Haloferacaceae Phosphoglycerate_dehydrogenase     5  1.72610114
2     Methanoperedenaceae Phosphoglycerate_dehydrogenase     1 -0.60015209
3            unclassified Phosphoglycerate_dehydrogenase     1 -0.60015209
551         Acidilobaceae Phosphoglycerate_dehydrogenase     1 -0.60015209
552          unclassified Phosphoglycerate_dehydrogenase     1 -0.60015209
567         Acidilobaceae Phosphoglycerate_dehydrogenase     1 -0.60015209
690      Archaeoglobaceae Phosphoglycerate_dehydrogenase     2 -0.01858878
14017       Haloferacaceae             Asparagine_synthase     2  0.08780760
14018 Methanoperedenaceae             Asparagine_synthase     3  0.89748609
14019        unclassified             Asparagine_synthase     0 -1.53154939
```

```
14567       Acidilobaceae            Asparagine_synthase     1 -0.72187090
14568        unclassified            Asparagine_synthase     1 -0.72187090
14583       Acidilobaceae            Asparagine_synthase     1 -0.72187090
14706    Archaeoglobaceae            Asparagine_synthase     1 -0.72187090
63073      Haloferacaceae          Alanine_dehydrogenase     5  1.53589524
63074 Methanoperedenaceae          Alanine_dehydrogenase     1 -0.26825391
63075        unclassified          Alanine_dehydrogenase     0 -0.71929120
63623       Acidilobaceae          Alanine_dehydrogenase     0 -0.71929120
63624        unclassified          Alanine_dehydrogenase     1 -0.26825391
63639       Acidilobaceae          Alanine_dehydrogenase     0 -0.71929120
63762    Archaeoglobaceae          Alanine_dehydrogenase     0 -0.71929120
68329      Haloferacaceae      Isopropyl_malate_synthase     3  0.49494473
68330 Methanoperedenaceae      Isopropyl_malate_synthase     4  1.23231138
68331        unclassified      Isopropyl_malate_synthase     1 -0.97978855
68879       Acidilobaceae      Isopropyl_malate_synthase     0 -1.71715520
68880        unclassified      Isopropyl_malate_synthase     2 -0.24242191
68895       Acidilobaceae      Isopropyl_malate_synthase     0 -1.71715520
69018    Archaeoglobaceae      Isopropyl_malate_synthase     2 -0.24242191
```

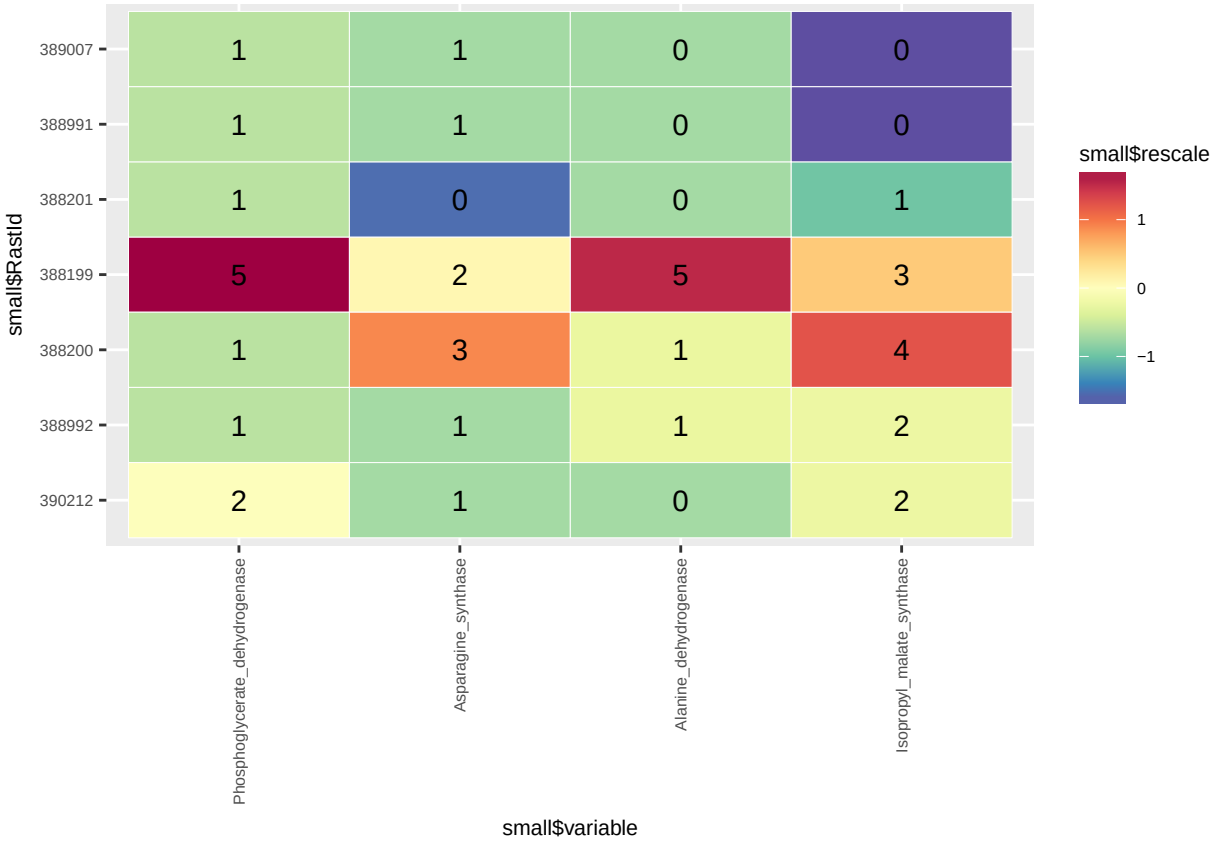
```
##Bueno aqui voy
##Heat.expansion<-adply(HP_Archaeas_Taxa.m[!HP_Archaeas_Taxa.m$variable %in% c("Co


#########################################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)))+ geo
```




```
## Graph coloured by columns, showing true values, not scaled
ggplot(small, aes(small$variable, small$RastId,label=round(small$value,digits=1)))
```




```
#########################################################
### Color according to mean plus standar deviation
####

################## trying to make heatplot colors on dataframe

small.m<- ddply(small, .(variable),summarize, mean = round(mean(value), 2),sd = round(s
rownames(small.m)<-small.m$variable

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

Chapter 2

Archaea EvoMining Results

During the decade between 1970 and 1980, Archaea was recognized as new life domain, a kingdom different from Bacteria and Eucarya in an exciting first great application of 16S phylogeny[112] . Main differences between this kingdoms are that Archaeal DNA is not arranged in a nucleus as in Eucarya and Archaeal celular walls are not composed from peptidoglycans as in Bacteria. Archaeal proteins may be higlhy valuable to biotechnology industry for their great stability due to extreme temperature, PH and salt content conditions on Archeal habitats. Despite no Archaeal Natural products biosynthetic gene clusters (BGC's) has been reported on MiBIG, Archaea do have BGC's, some of them seems to be acquired by horizontal gene transfer (HGT) like methano nrps {search reference}. Other Archeal natural products known are archaeosins, Diketopiperazines, Acyl Homoserine Lactones, Exopolysaccharides, Carotenoids, Biosurfactants, Phenazines and Organic Solutes but this knowledge is not comparable to Bacterial BGC's knowledge[96].

Natural products biosynthetic gene clusters search is actually performed using either *high-confidence/low-novelty or low-confidence/high-novelty* bioinformatic approaches [47]. 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 [62]. Also on Actinobacteria antiSMASH analysis on 1245 genomes found 774 different classes of natural products, the same analysis on 876 Archaeal genomes, a full kingdom, identifies only 35 BGC's classes. So either Archaea does not have natural products BGC's or this are not yet known. Next paragraph deals with a possible approach about how natural products BGC's can be find.

Archaea resembled Bacteria in that Archaea uses horizontal gene transfer as a genic interchange mecanism, Archaeal genomes contains operons [115] 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 [115]. General lack of introns allows automatic genome annotation, operons gene organization permits functional inference to a certain degree and HGT contribute to expansions on Archaeal genomes. Some phylum on Archaea has an open pangenome, and as we will show on this chapter some Archaea has central pathway expansions. Enzyme families from central pathways expansions, open pangenome and operon organization made EvoMining succesful on Actinobacteria, this lead us to think that evoMining is suitable to analize Archaeal genomes, even more since EvoMining is a method oriented to use evolution and its not entirelyy based on previous knowledge of BGC's sequences if evolutionary logic behave on Archaea as on bacteria, new BGC's classes may be be found on Archaea.

EvoMining is a trade off between conserved known central metabolic function and enough expansions divergence on sequence and on clusters to divergence

2.1 Tables

Table 2.1: 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.

2.1.1 Expansions BoxPlot by metabolic family

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

Here is a reference to the expansion boxplot: Figure 2.2.

Are g
ev

Abstract

On 1977 Carl Woese based on phylogeneti
as a separate kingdom from Bacteria. There
available at NCBI, where it is possib
expansions. Even though it is known Arch
products they are poorly detected by hom
mining approaches. Therefore Archaea

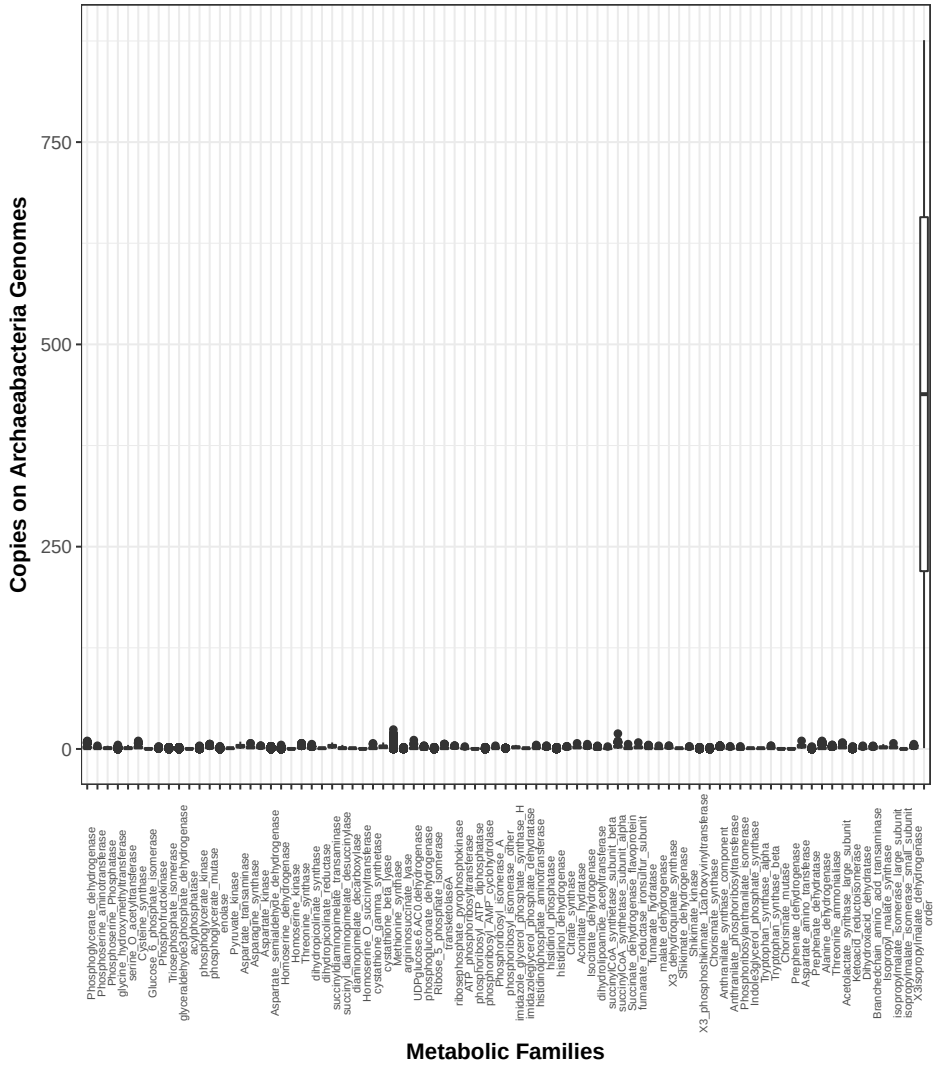
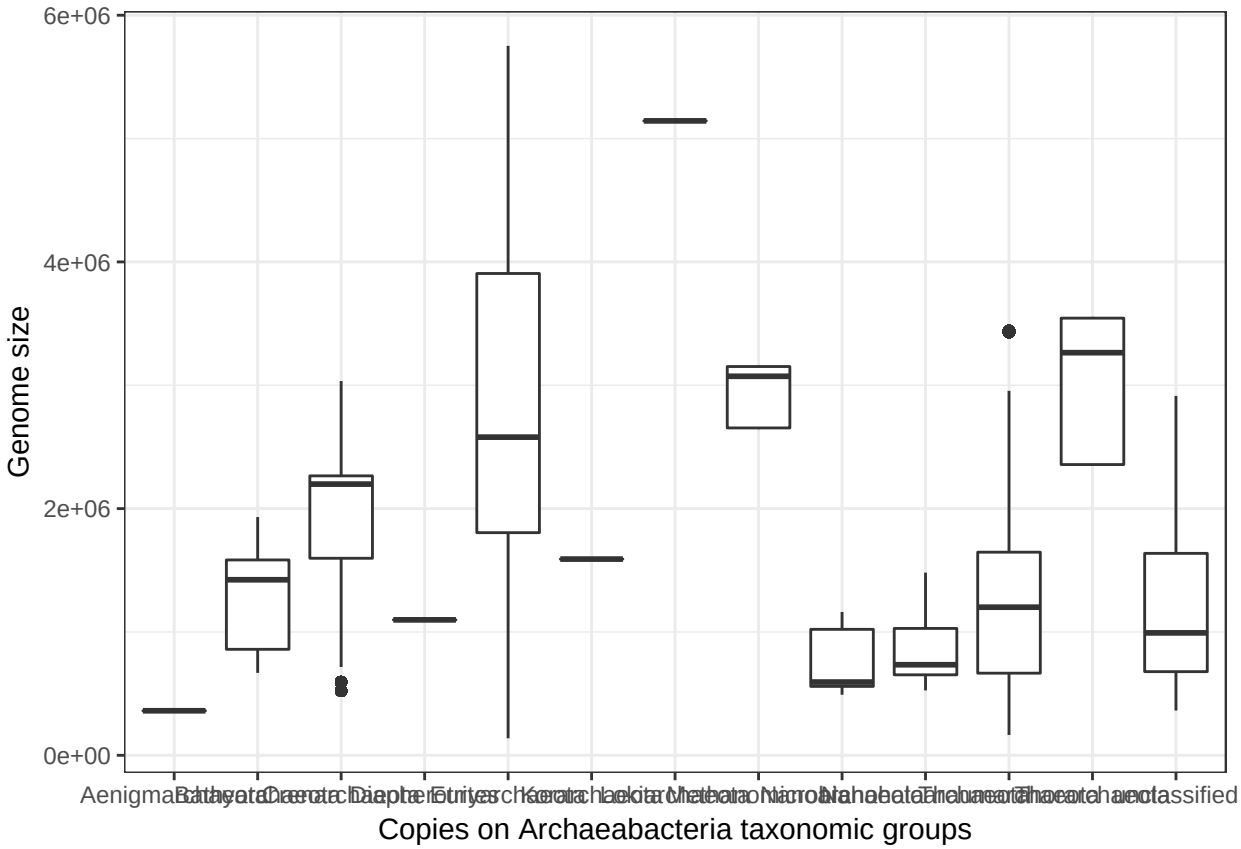


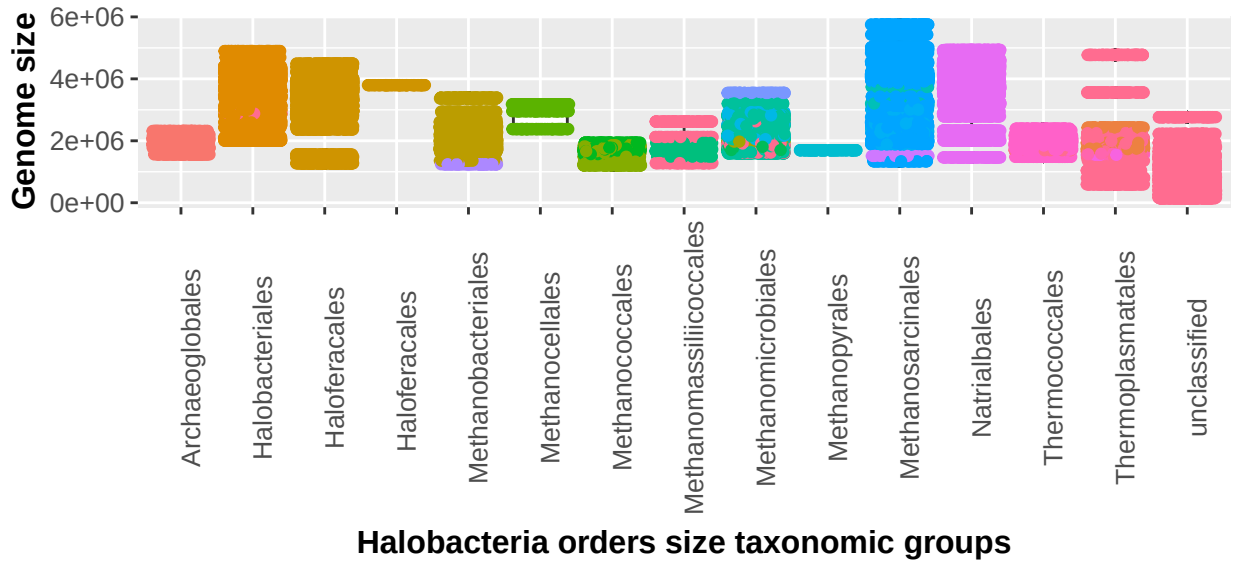
Figure 2.2: Expansions Boxplot

2.1.2 Expansions BoxPlot by metabolic family by phylum

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

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

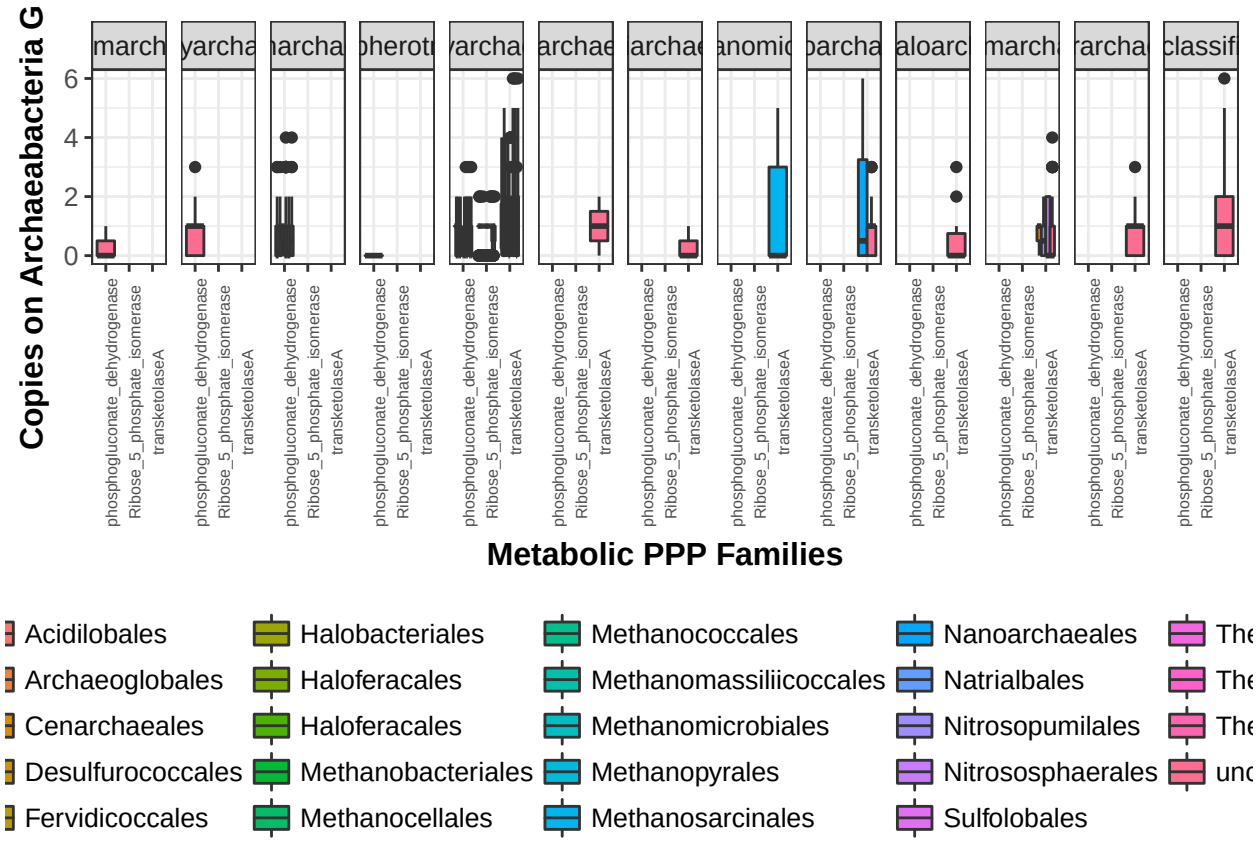



```
#+ geom_jitter(aes(color=ArchaeasHeatPlotBP.m$Phylum))


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

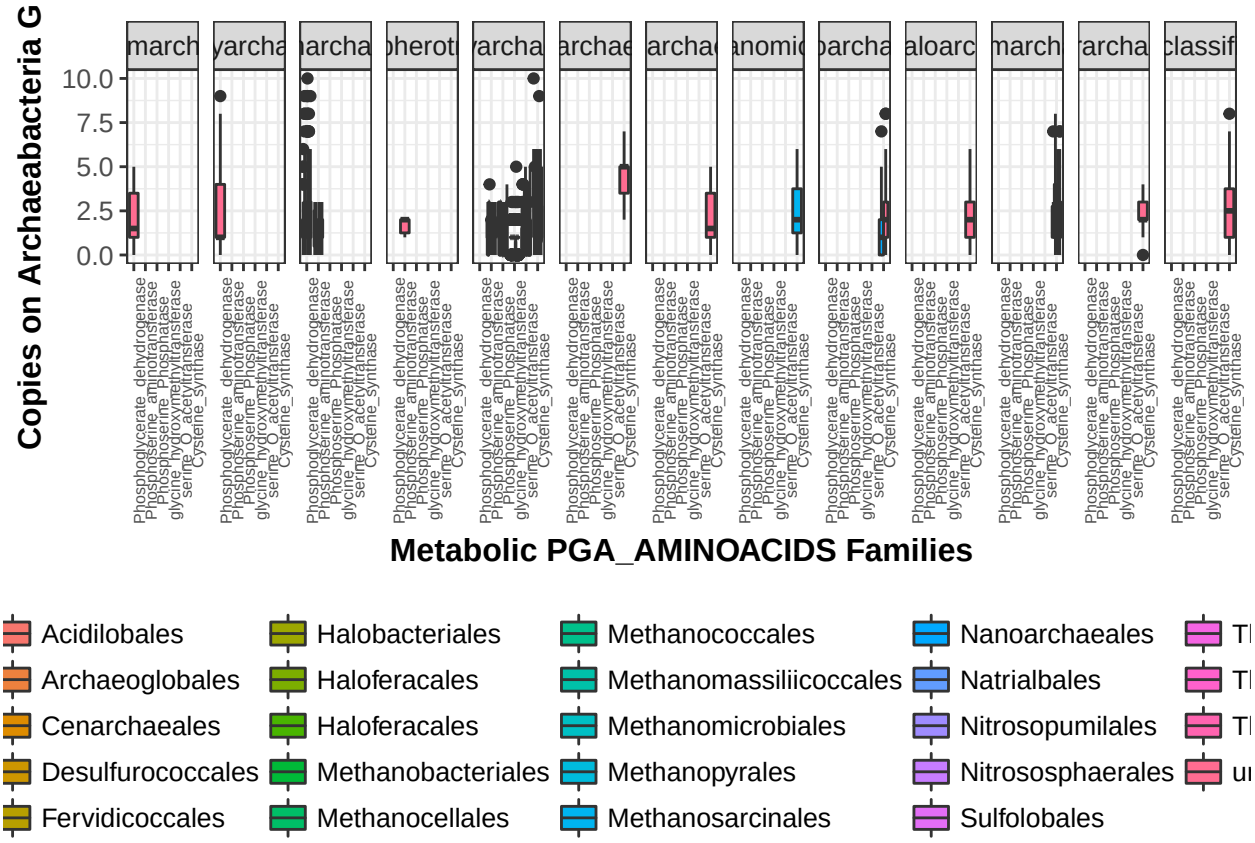


haeoglobaceae
roplasmaceae
obacteriaceae
oferacaceae
thanobacteriaceae
Methanocalculaceae
Methanocaldococcaceae
Methanocellaceae
Methanococcaceae
Methanocorpusculaceae
Methanomassiliicoccaceae
Methanomicrobiaceae
Methanoperedenaceae
Methanopyraceae
Methanoregulaceae
Methanosaetaceae
Methanosarcinaceae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

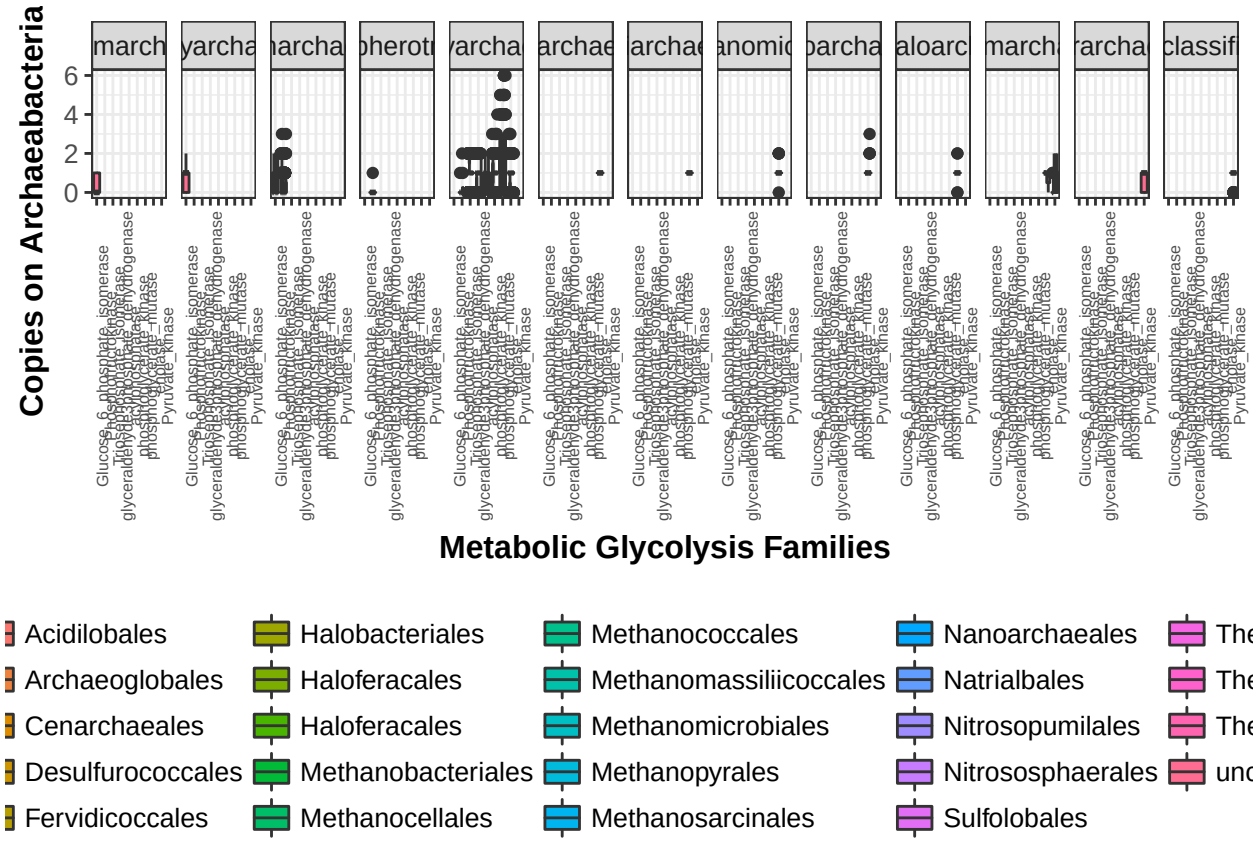
```
#MetFam_BP.m=subset(ArchaeasHeatPlotBP.m,Family=="Methanosarcinaceae")
#ggplot(MetFam_BP.m, aes(x=MetFam_BP.m$Size, y=MetFam_BP.m$value))
#+theme(plot.title = element_text(size = 14, face = "bold"), text = element_text(size


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


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



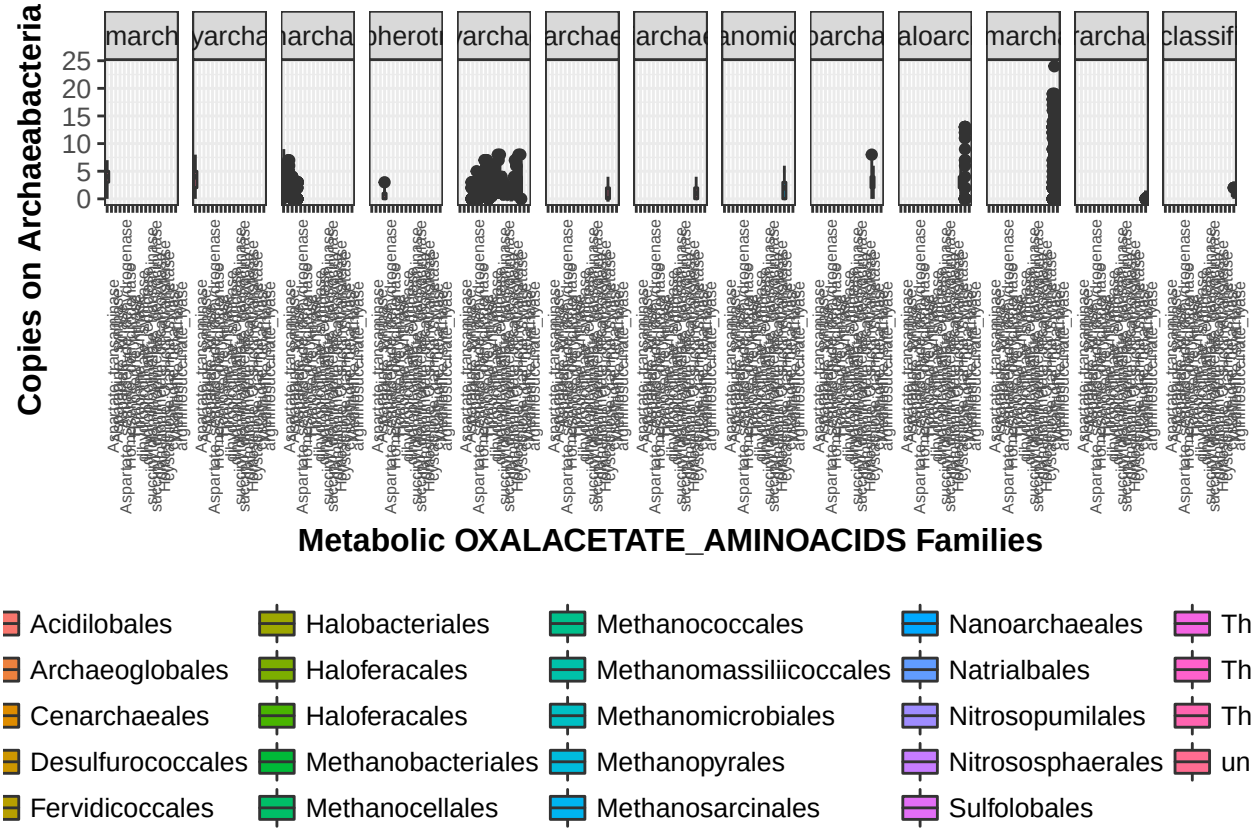

```
#+ geom_jitter(aes(color=MetFam_BP.m$Phylum))

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



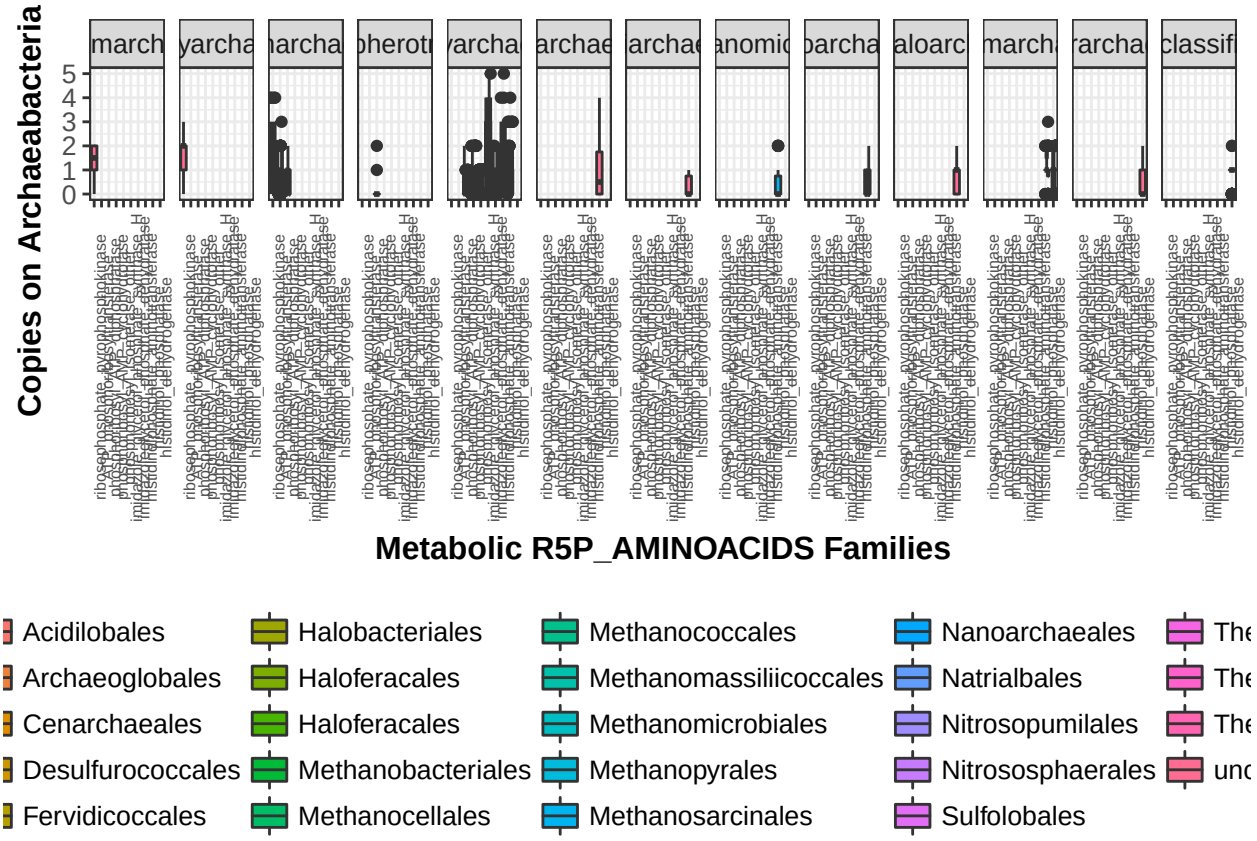

```
#+ geom_jitter(aes(color=MetFam_BP.m$Phylum))

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
```



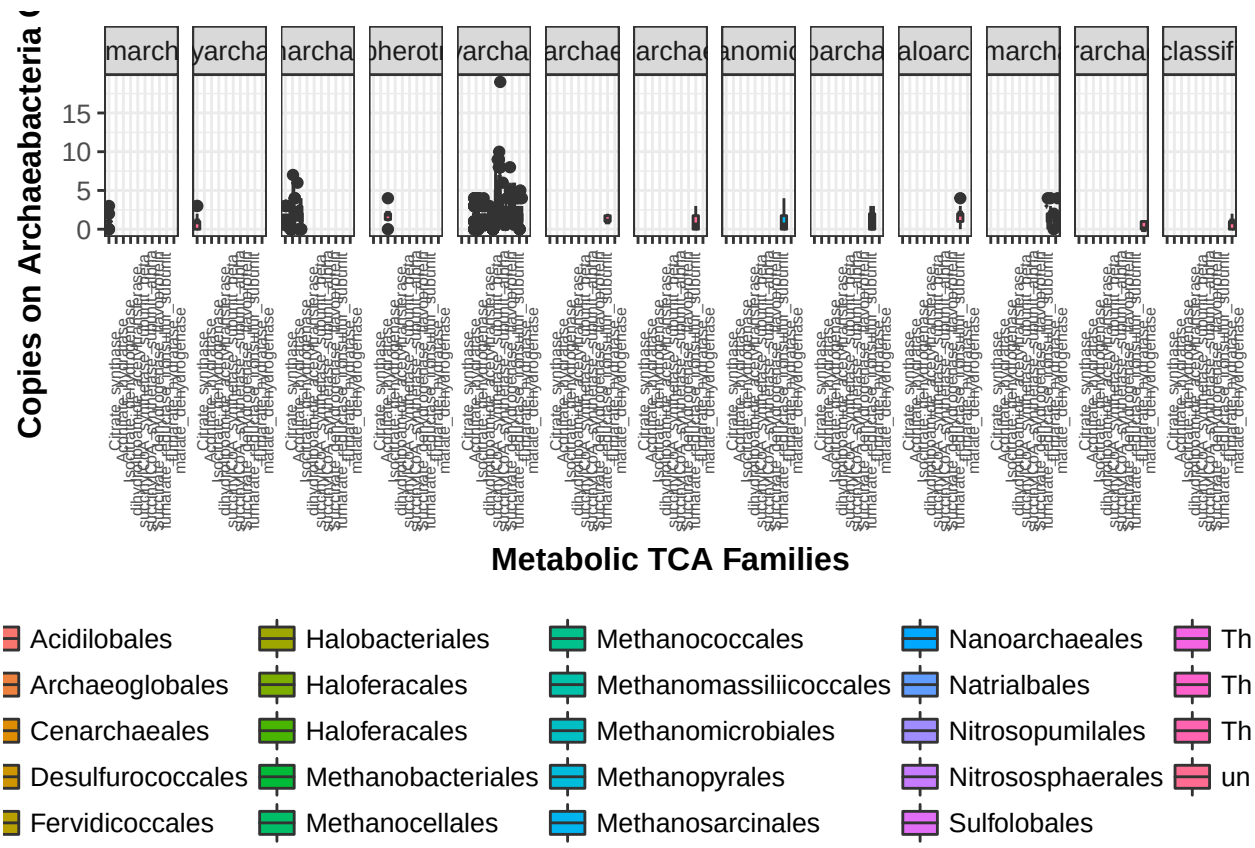

```
#+ geom_jitter(aes(color=MetFam_BP.m$Phylum))

MetFam=subset(ArchaeasCentral,Pathway=="OXALACETATE_AMINOACIDS")
MetFam_BP.m=ArchaeasHeatPlotBP.m[ArchaeasHeatPlotBP.m$variable %in% MetFam$Enzyme,
ggplot(MetFam_BP.m, aes(x=MetFam_BP.m$variable, y=MetFam_BP.m$value, fill=Order))+
```



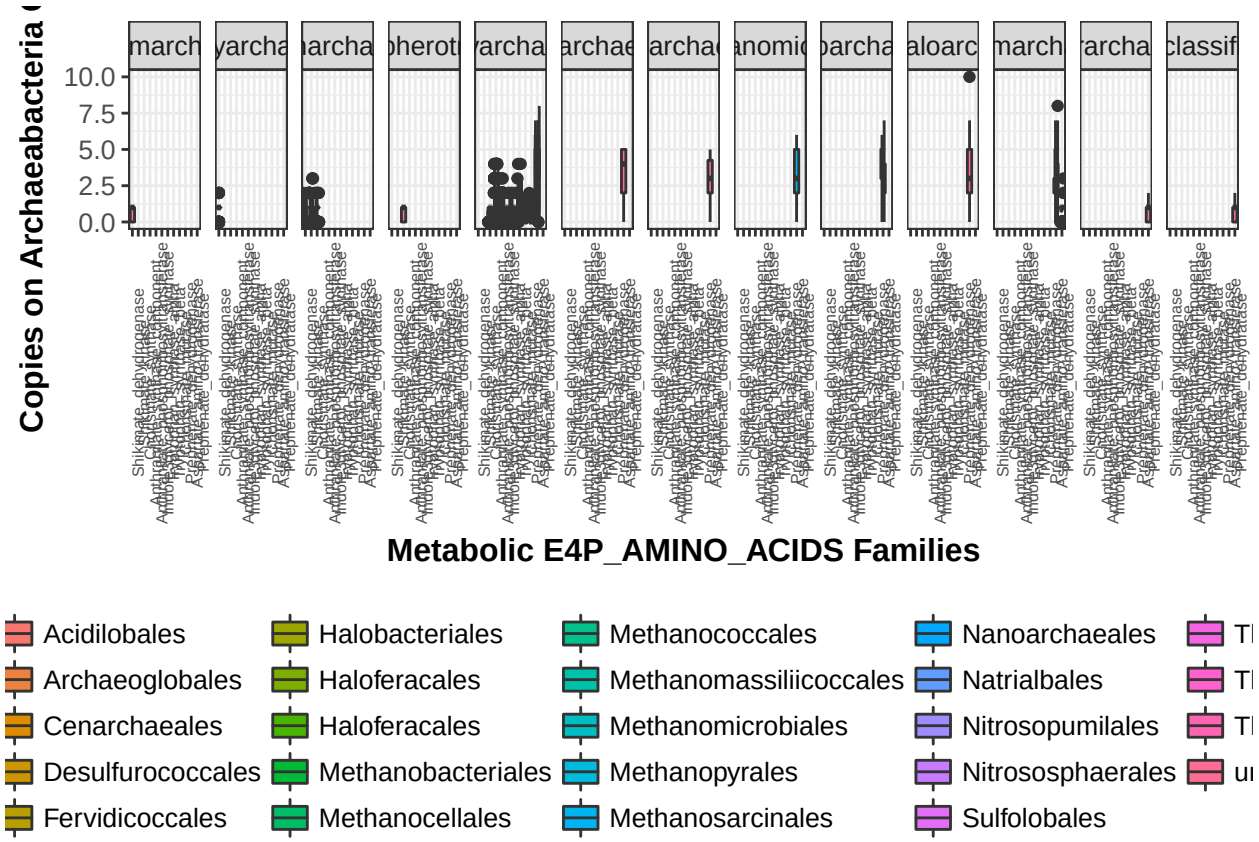

```
#+ geom_jitter(aes(color=MetFam_BP.m$Phylum))

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



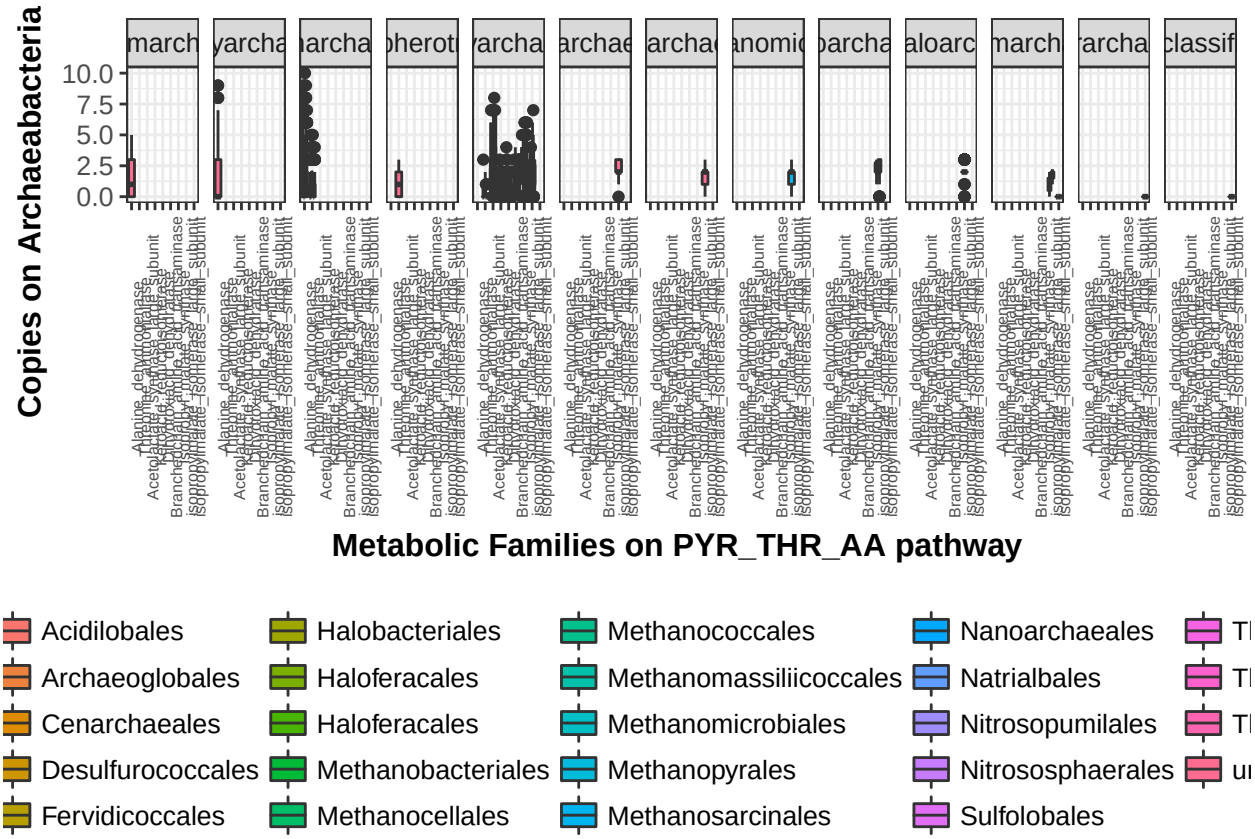

```
 #+ geom_jitter(aes(color=MetFam_BP.m$Phylum))

#
MetFam=subset(ArchaeasCentral,Pathway=="TCA")
MetFam_BP.m=ArchaeasHeatPlotBP.m[ArchaeasHeatPlotBP.m$variable %in% MetFam$Enzyme,
ggplot(MetFam_BP.m, aes(x=MetFam_BP.m$variable, y=MetFam_BP.m$value, fill=Order))+
```




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

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



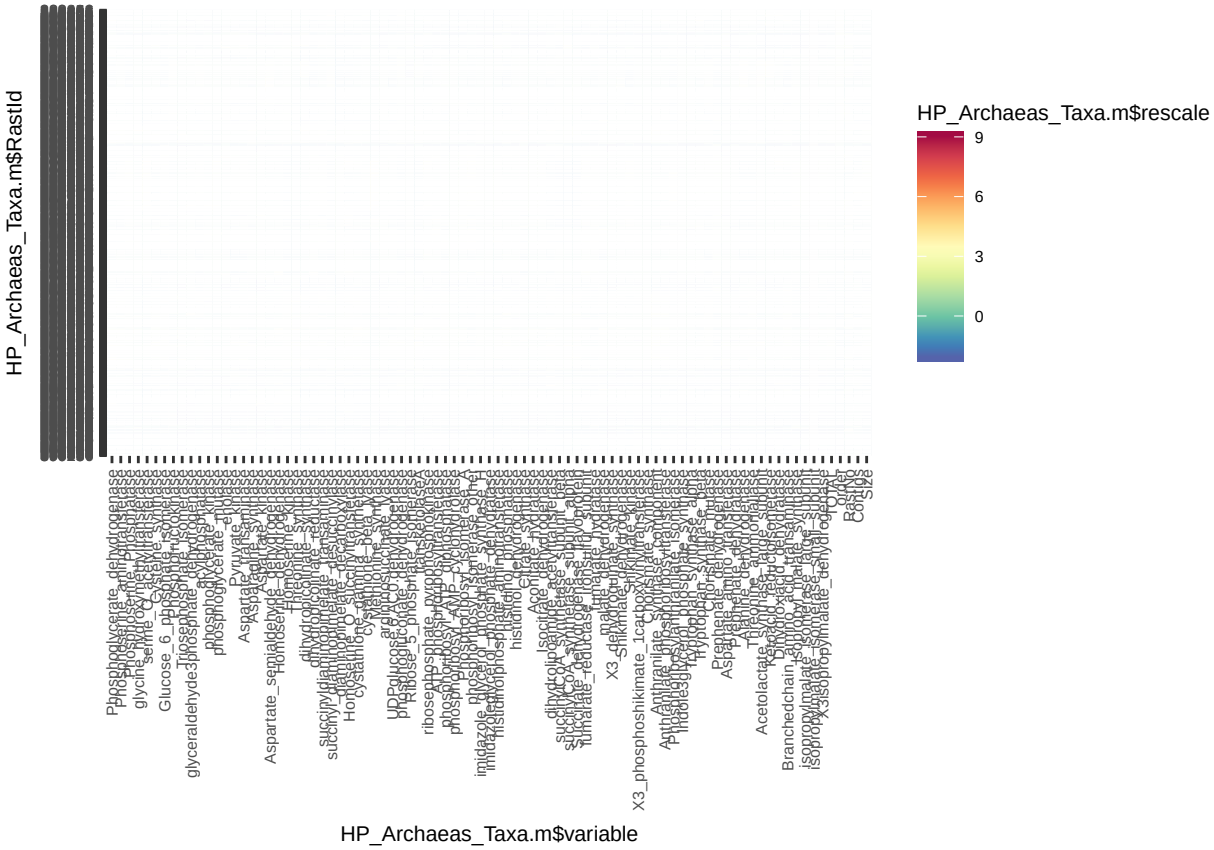

```
#+ geom_jitter(aes(color=MetFam_BP.m$Phylum))


MetFam=subset(ArchaeasCentral,Pathway=="PYR_THR_AA")
MetFam_BP.m=ArchaeasHeatPlotBP.m[ArchaeasHeatPlotBP.m$variable %in% MetFam$Enzyme,
ggplot(MetFam_BP.m, aes(x=MetFam_BP.m$variable, y=MetFam_BP.m$value, fill=Order))+
```




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

#ggsave("chapter3/expansion_plotArchaeas.pdf", plot = expansion_plotArchaea,height = 8
```

2.2 Central pathway expansions

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



Here is a reference to the HeatPlot: Figure 2.3.

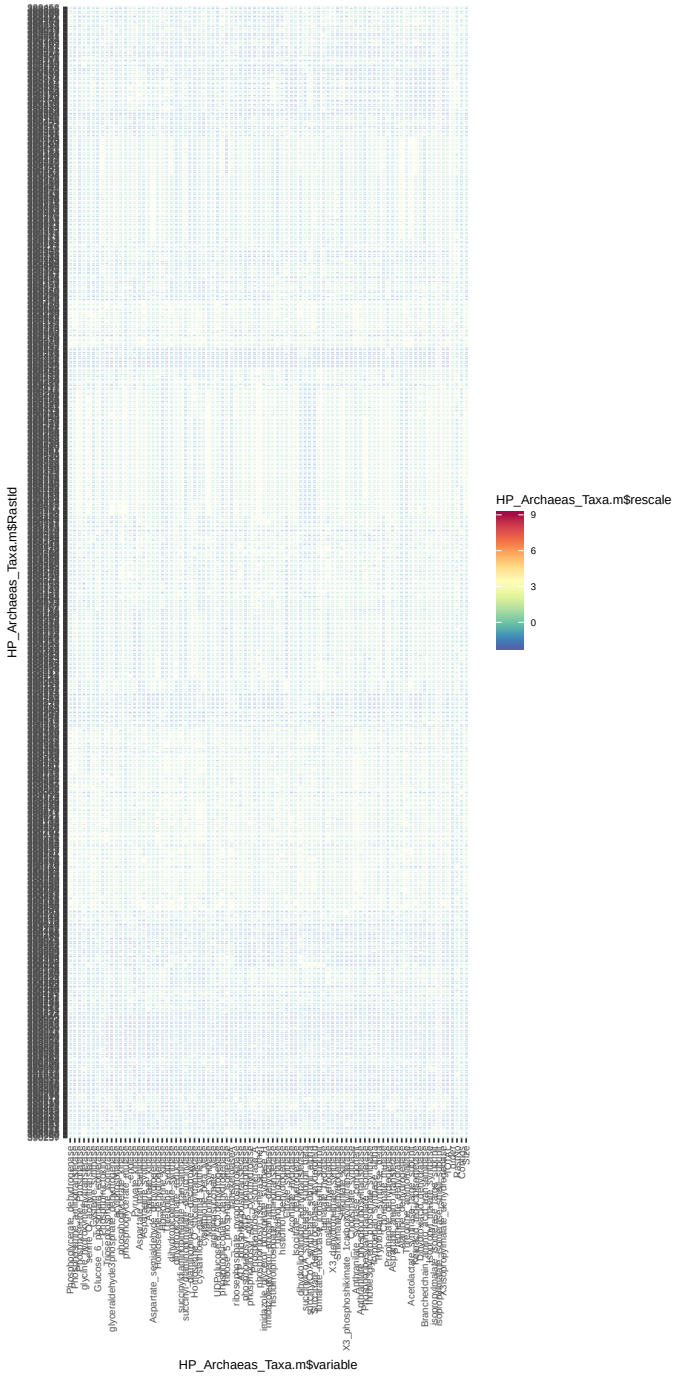


Figure 2.3: Archaeas Heatplot

2.3 Genome Size correlations

2.3.1 Correlation between genome size and AntiSMASH products

Genome size vs Total antismash cluster coloured by order

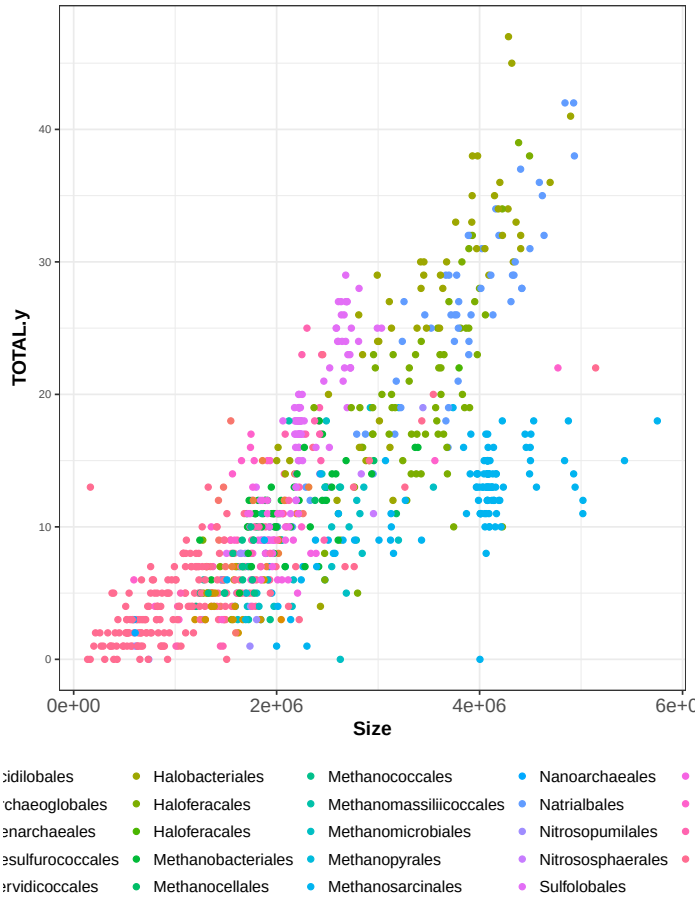


Figure 2.4: Correlation between Archaeas genome size and antismash Natural products detection colored by Order

Here is a reference to Genome size vs Total antismash cluster: Figure 2.4.

Genome size vs Total antismash cluster detected splitted by order

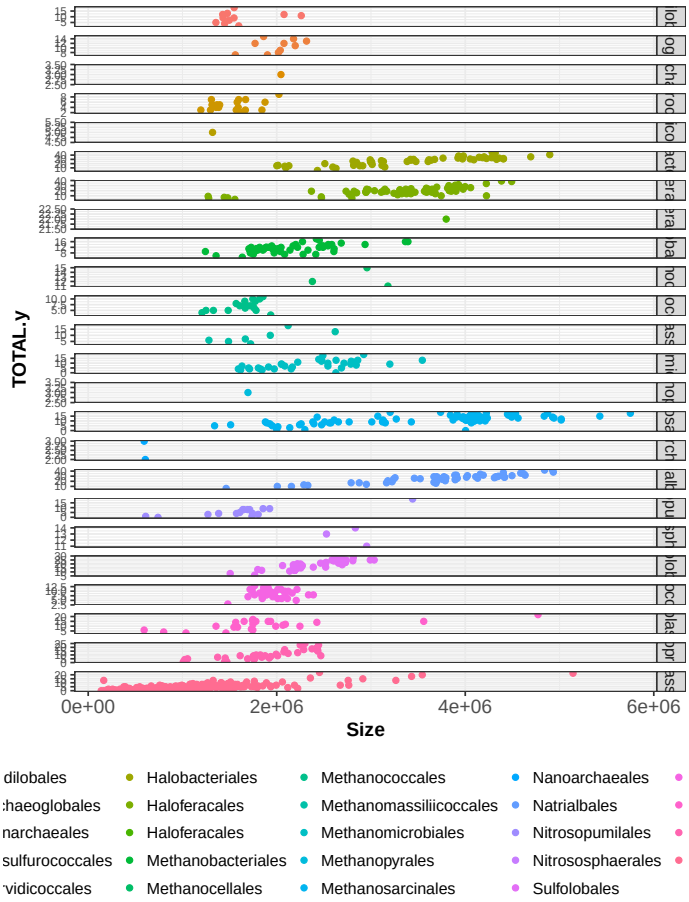


Figure 2.5: Correlation between Archaeas genome size and antismash Natural products detection grided by Order

Here is a reference to Correlation between genome size and antismash Natural products detection grided by Order plot: Figure 2.5.

2.3.2 Correlation between genome size and Central pathway expansions

Genome size vs Total central pathway expansion coloured by order

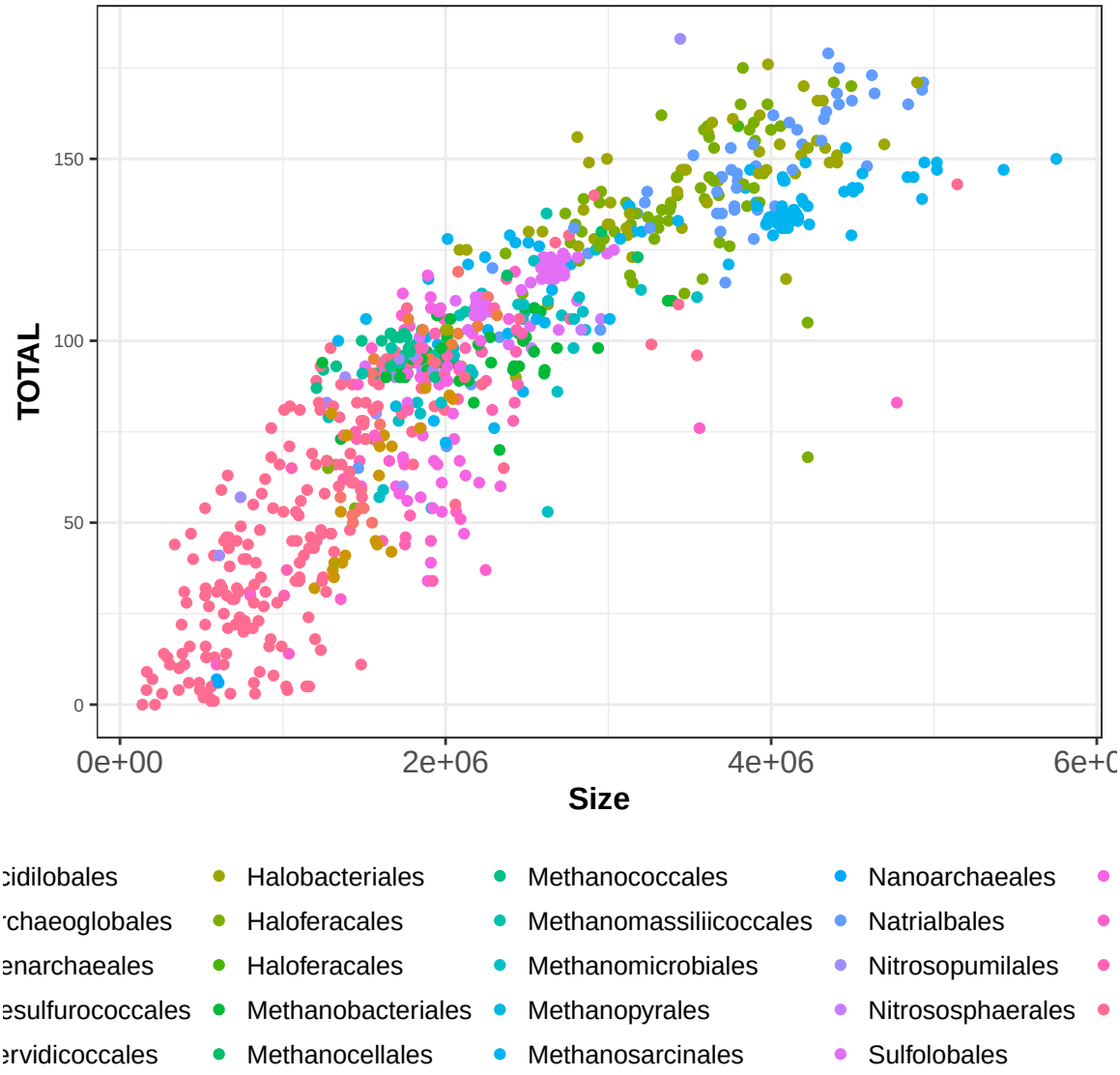


Figure 2.6: Correlation between Archaeas genome size and central pathway expansions

Here is a reference to the size vs Total central pathway expansion plot: Figure 2.6.

Genome size vs Total central pathway expansion grided by order

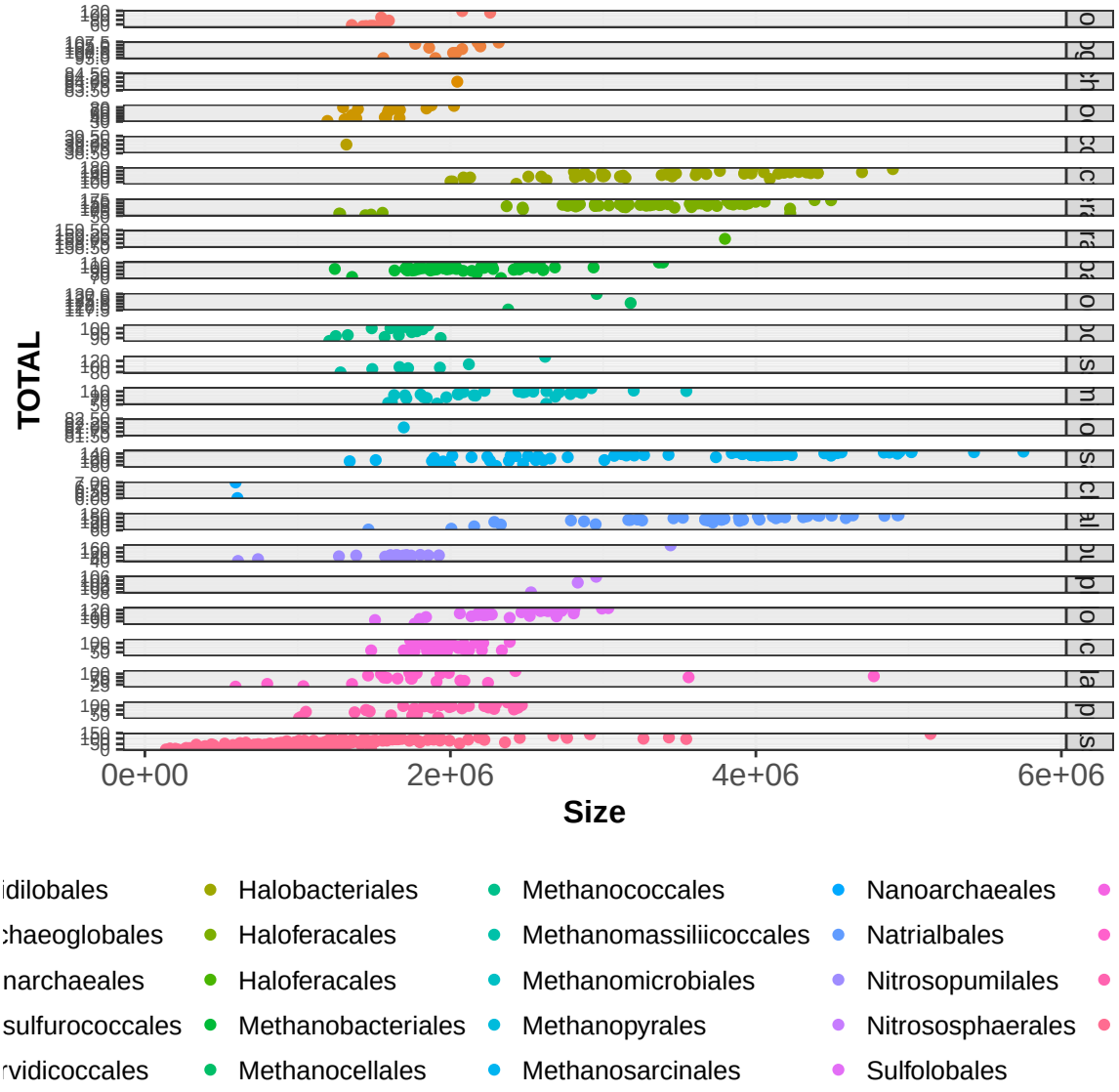


Figure 2.7: Correlation between Archaeas genome size and central pathway expansions grided by order

Here is a reference to the Genome size vs Total central pathway expansion grided by order plot: Figure 2.7.

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

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

```
Warning: Removed 65604 rows containing missing values (geom_point).
```

Genome size vs Total central pathway expansion coloured by metabolic Family

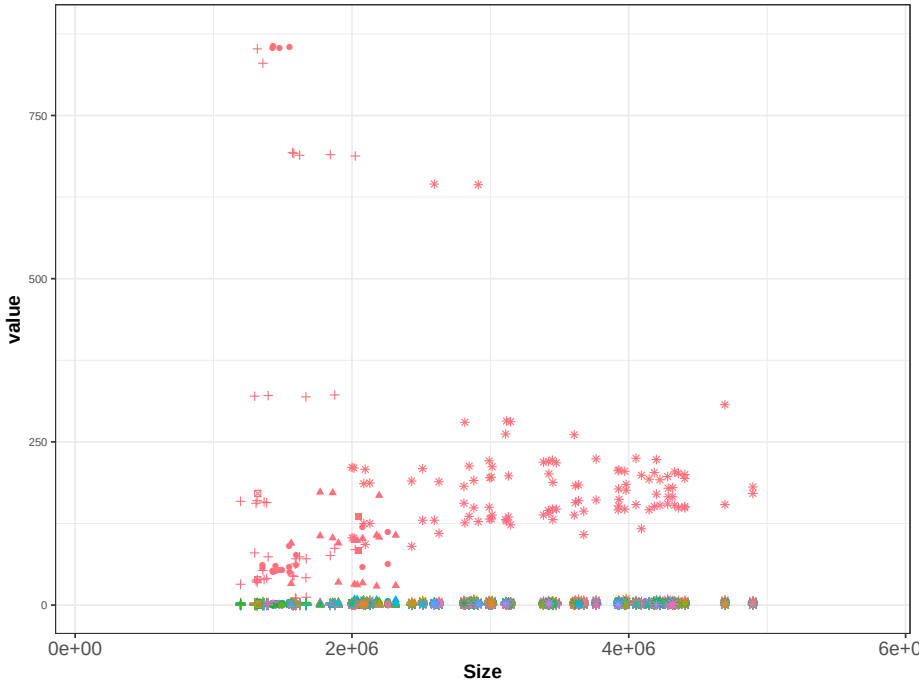


Figure 2.8: Correlation between Archaeas Genome size vs Total central pathway expansion coloured by metabolic Family

Here is a reference to the Genome size vs Total central pathway expansion coloured by metabolic Family plot: Figure 2.8.

Future Work: Genome size vs Total central pathway expansion grided by metabolic Family For clarity I need to also grid and group by Metabolic Pathway

Here is a reference to Genome size vs Total central pathway expansion grided by metabolic Family plot: **??**.

2.4 Natural products

2.4.1 Natural products recruitments from EvoMining heatplot

We can see natural products recruitment after central pathways expansions colored by their kingdom.
Natural products recruited by metabolic family, colored by phylogenetic origin.

Recruitments after central pathways expansions coloured by Kingdom

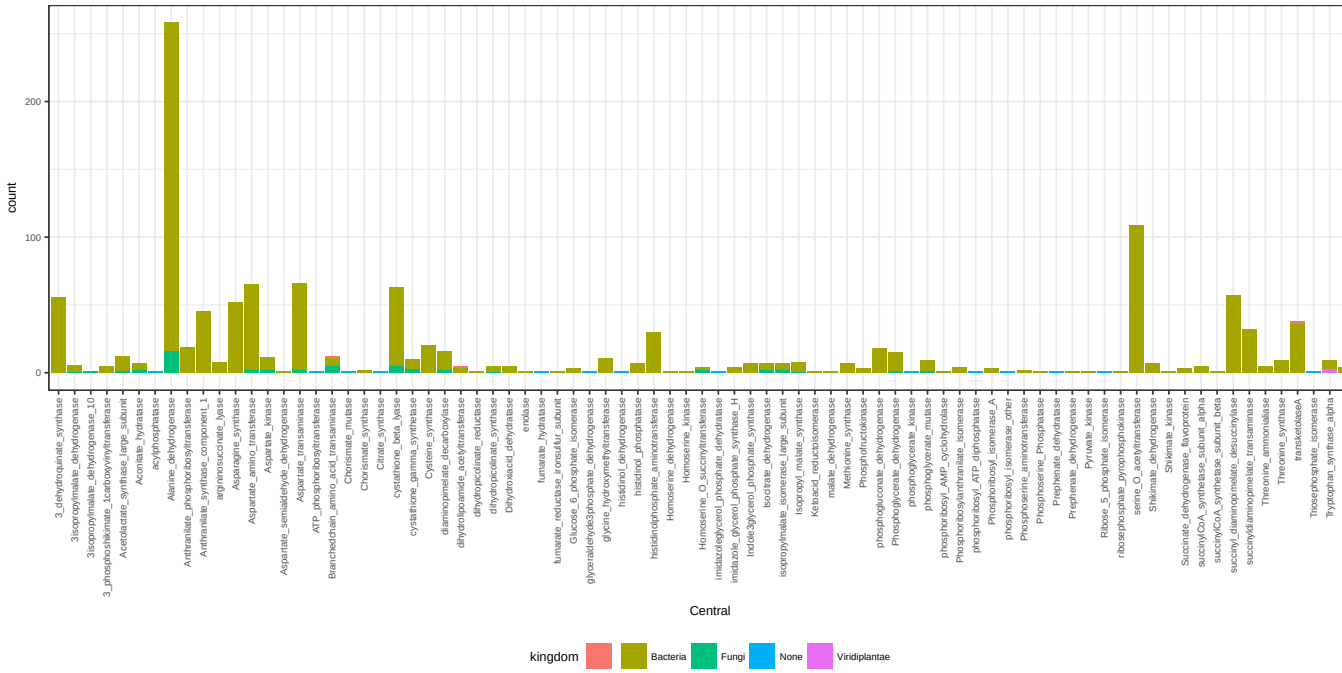


Figure 2.9: Archaeas Recruitmens on central families coloured by kingdom

Here is a reference to Recruitments after central pathways expansions colourd by Kingdom plot: Figure 2.9.

Recruitments after central pathways expansions colourd by taxonomy

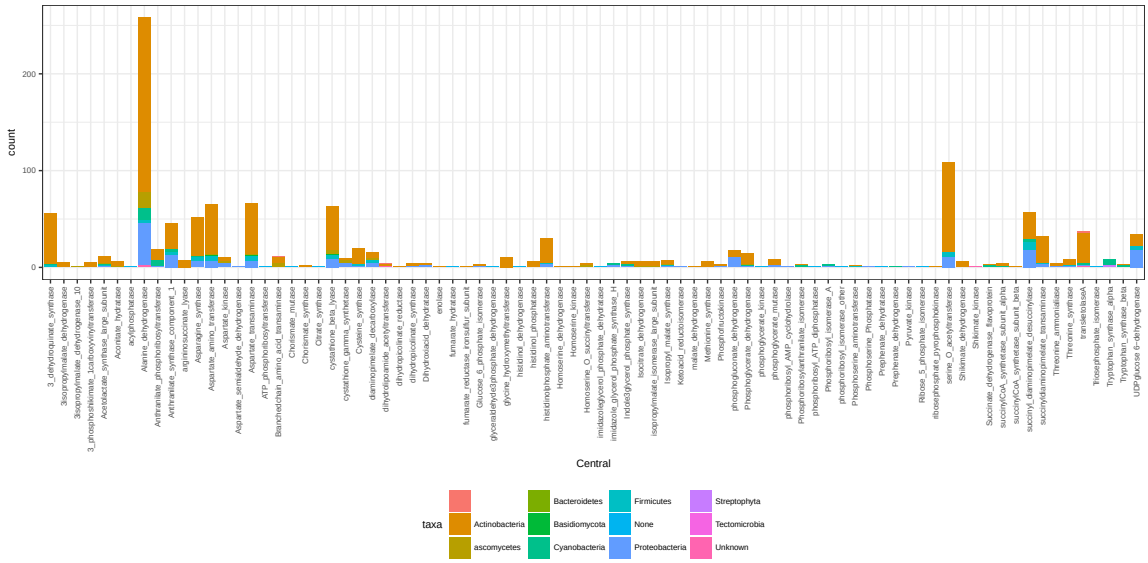


Figure 2.10: Archaeas Recruitmens on central families coloured by taxonomy

Here is a reference to Recruitments after central pathways expansions colourd by taxa plot: Figure 2.10.

2.5 Archaeas AntiSMASH

Taxonomical diversity on Archaeasbacteria Data

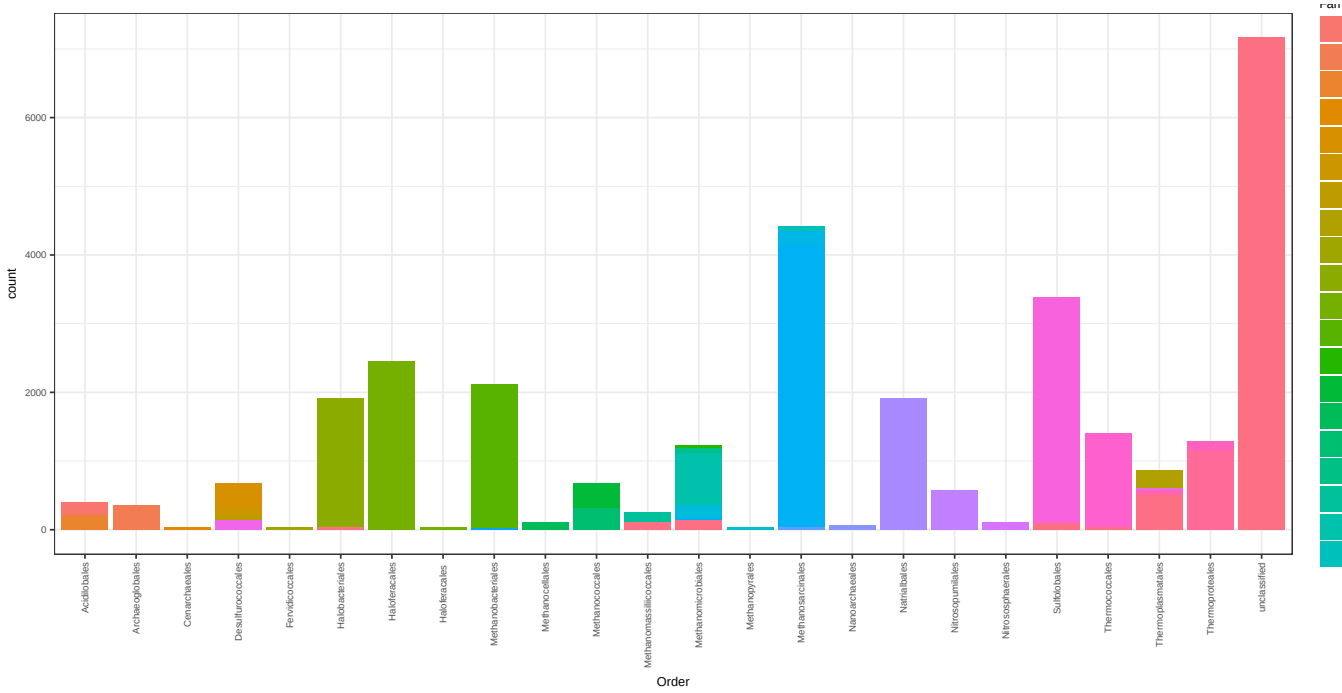


Figure 2.11: Archaeas Diversity

Here is a reference to Recruitments after central pathways expansions colourd by taxa plot: Figure 2.11.

Smash diversity

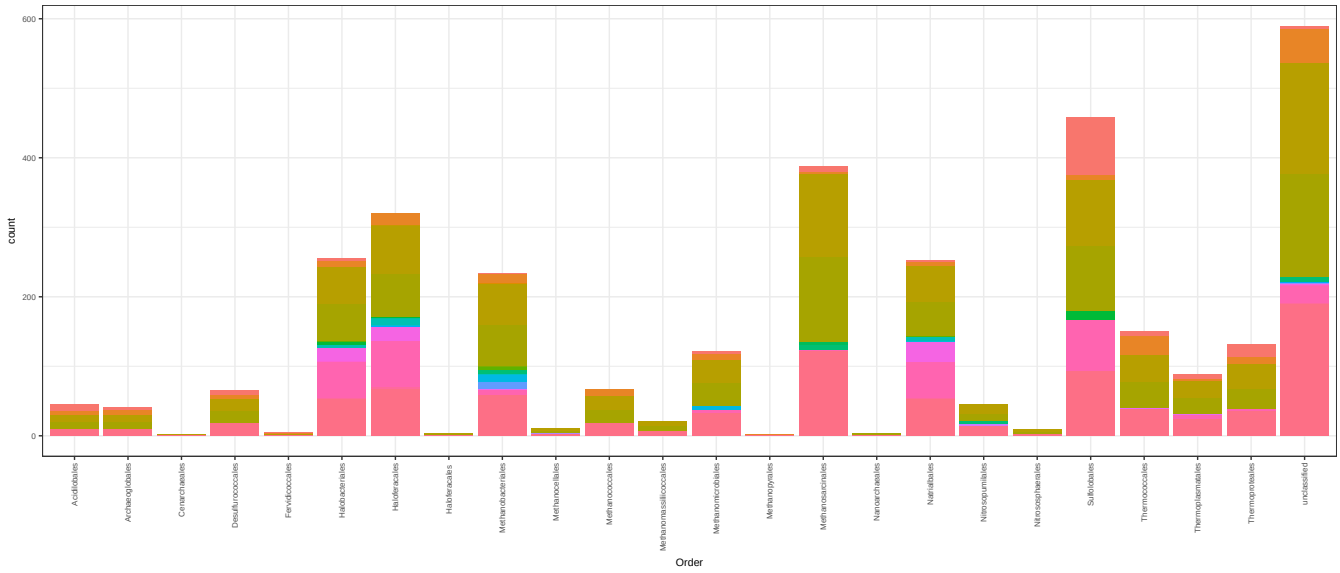


Figure 2.12: Archaeas Smash Taxonomical Diversity

Here is a reference to Recruitments after central pathways expansions colourd by taxa plot: Figure 2.12.

2.5.1 AntisSMASH vs Central Expansions

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

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

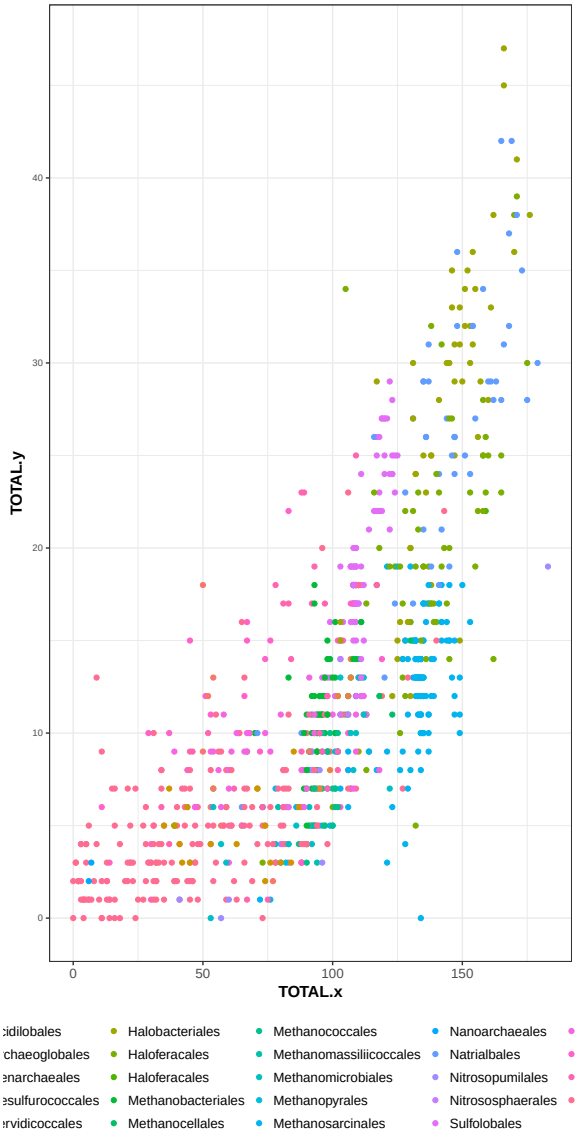


Figure 2.13: Correlation between Archaeas central pathway expansions and antismash Natural products detection

Here is a reference to the expansions vs antismash NP's clusters plot: Figure 2.13.

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

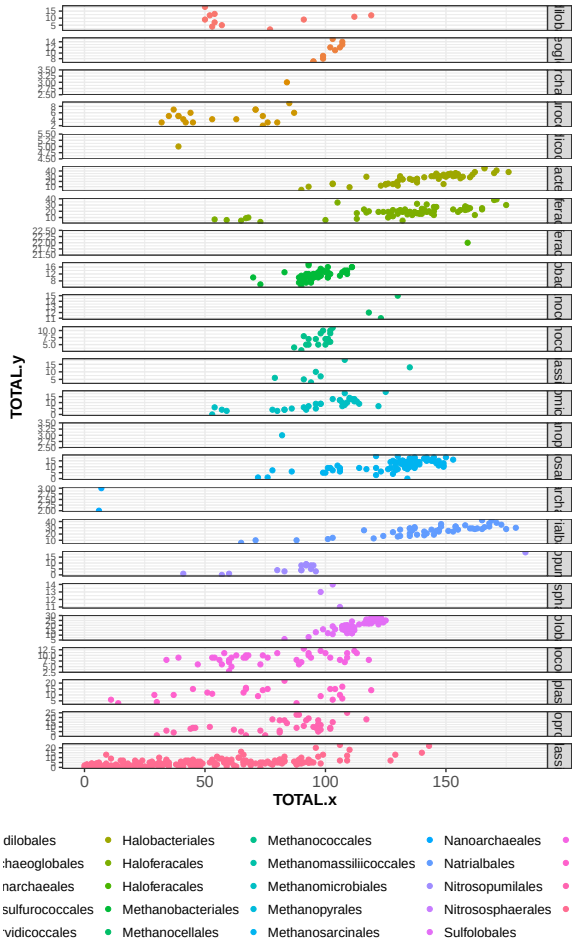


Figure 2.14: Correlation between Archaeas central pathway expnasions and antismash Natural products detection

Here is a reference to the expansions vs antismash NP's clusters splitted by order plot Figure 2.14.

AntisMAsh vs Expansions by taxonomic Family

Natural products colured by family

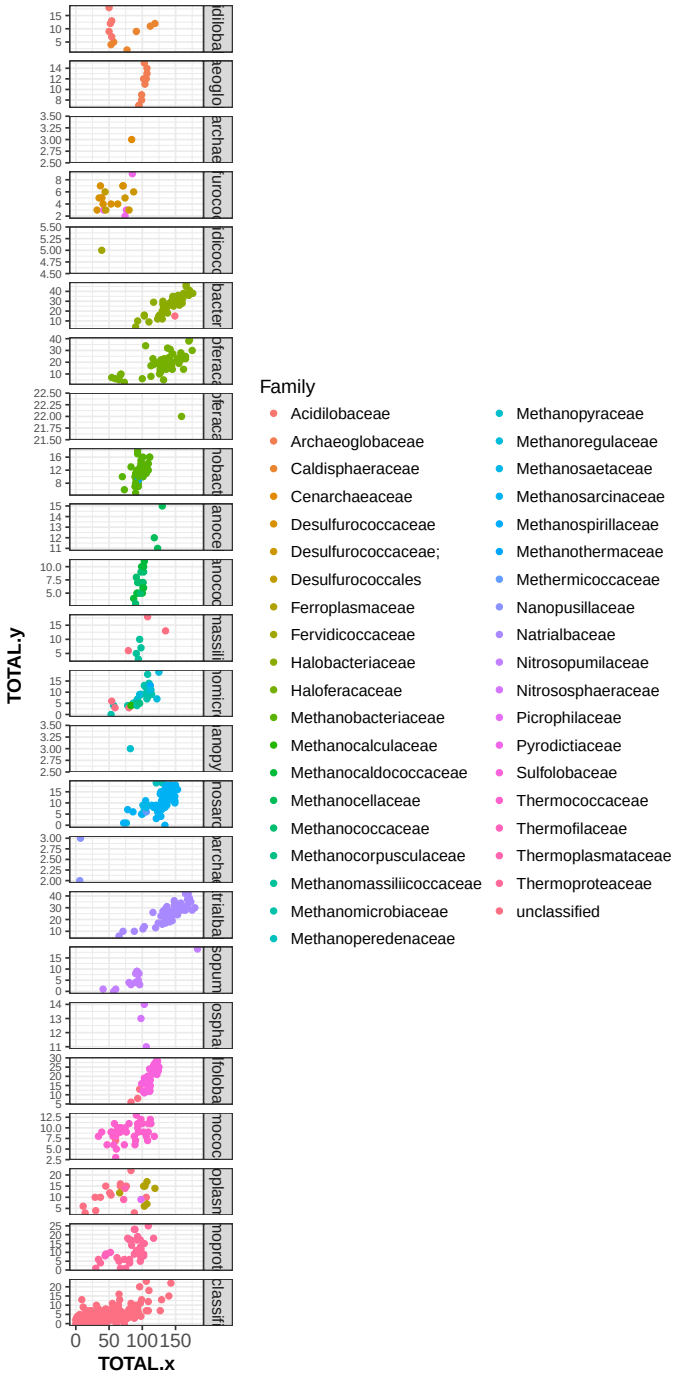


Figure 2.15: Archaeas Natural products by family

Here is a reference to the Natural products colured by family plot Figure 2.15.

2.6 Selected trees from EvoMining

Phosphoribosyl_isomerase_3 family
Figure from EvoMining

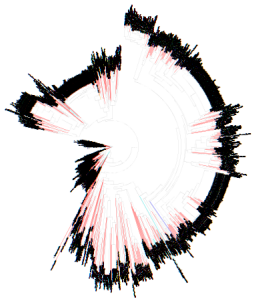


Figure 2.16: Phosphoribosyl isomerase A EvoMiningtree

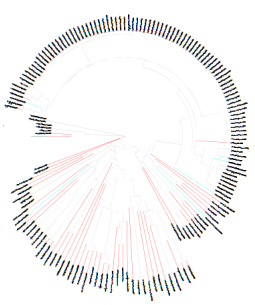


Figure 2.17: Phosphoribosyl isomerase other EvoMiningtree

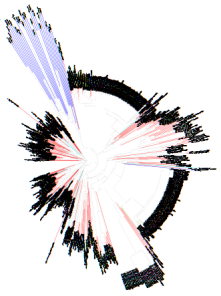


Figure 2.18: Phosphoribosyl anthranilate isomerase EvoMiningtree

2.7

Other possible databases Archaeal signatures *set of protein-encoding genes that function uniquely within the Archaea; most signature proteins have no recognizable bacterial or eukaryal homologs* [114] ## Footnotes and Endnotes

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

2.8 Bibliographies

Of course you will need to cite things, and you will probably accumulate an armful of sources. There are a variety of tools available for creating a bibliography database (stored with the .bib extension). In addition to BibTeX suggested below, you may want to consider using the free and easy-to-use tool called Zotero. The Reed librarians have created Zotero documentation at `http://libguides.reed.edu/citation/zotero`. In addition, a tutorial is available from Middlebury College at `http://sites.middlebury.edu/zoteromiddlebury/`.

R Markdown uses *pandoc* (`http://pandoc.org/`) to build its bibliographies. One nice caveat of this is that you won't have to do a second compile to load in references as standard LaTeX requires. To cite references in your thesis (after creating your bibliography database), place the reference name inside square brackets and precede it by the "at" symbol. For example, here's a reference to a book about worrying: [137]. This `Molina1994` entry appears in a file called `thesis.bib` in the `bib` folder. This bibliography database file was created by a program called BibTeX. You can call this file something else if you like (look at the YAML header in the main .Rmd file) and, by default, is to placed in the `bib` folder.

For more information about BibTeX and bibliographies, see our CUS site (`http://web.reed.edu/cis/help/latex/index.html`)[2]. There are three pages on this topic: *bibtex* (which talks about using BibTeX, at `http://web.reed.edu/cis/help/latex/bibtex.html`), *bibtexstyles* (about how to find and use the bibliography style that best suits your needs, at `http://web.reed.edu/cis/help/latex/bibtexstyles.html`) and *bibman* (which covers how to make and maintain a bibliography by hand, without BibTeX, at `http://web.reed.edu/cis/help/latex/bibman.html`). The last page will not be useful unless you have only a few sources.

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

[1] footnote text
[2] [138]

can download a variety of different style files at `https://www.zotero.org/styles`. Make sure to download the file into the csl folder.

Tips for Bibliographies

- Like with thesis formatting, the sooner you start compiling your bibliography for something as large as thesis, the better. Typing in source after source is mind-numbing enough; do you really want to do it for hours on end in late April? Think of it as procrastination.
- The cite key (a citation's label) needs to be unique from the other entries.
- When you have more than one author or editor, you need to separate each author's name by the word "and" e.g. `Author = {Noble, Sam and Youngberg, Jessica},`.
- Bibliographies made using BibTeX (whether manually or using a manager) accept LaTeX markup, so you can italicize and add symbols as necessary.
- To force capitalization in an article title or where all lowercase is generally used, bracket the capital letter in curly braces.
- You can add a Reed Thesis citation[3] option. The best way to do this is to use the phdthesis type of citation, and use the optional "type" field to enter "Reed thesis" or "Undergraduate thesis."

2.9 Anything else?

If you'd like to see examples of other things in this template, please contact the Data @ Reed team (email `data@reed.edu`) with your suggestions. We love to see people using *R Markdown* for their theses, and are happy to help.

[3] [139]

Chapter 3

Actinobacteria EvoMining Results

Actinobacteria is an ancient phylum {Referencia de luis}

3.1 Tables

Table 3.1: Correlation of Inheritance Factors for Parents and Child

Factors	Correlation between Parents & Child
GenomeDB	1245
Families	65

3.1.1 Expansions BoxPlot by metabolic family

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

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

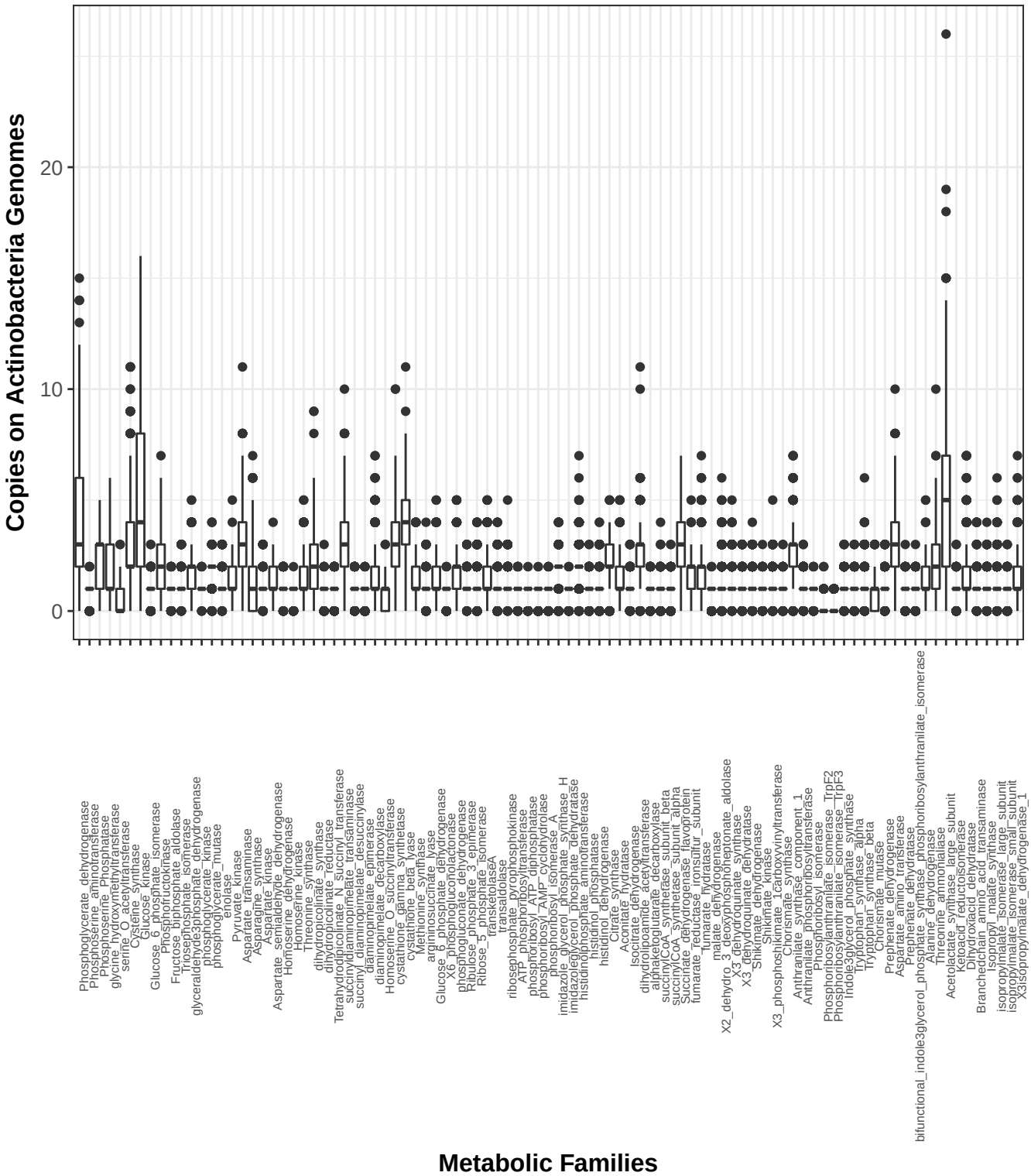
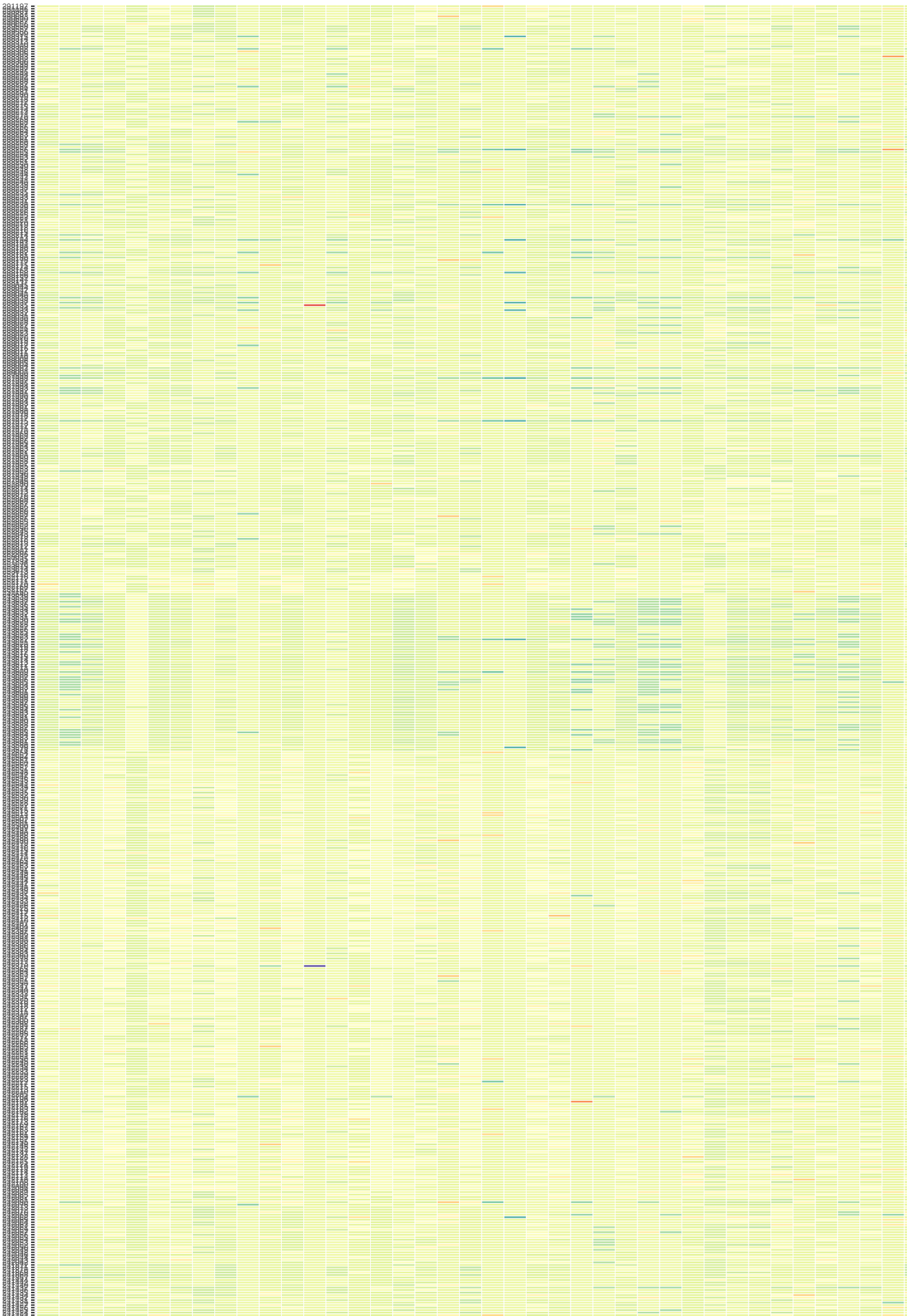


Figure 3.1: Expansions Boxplot

3.2 Central pathway expansions

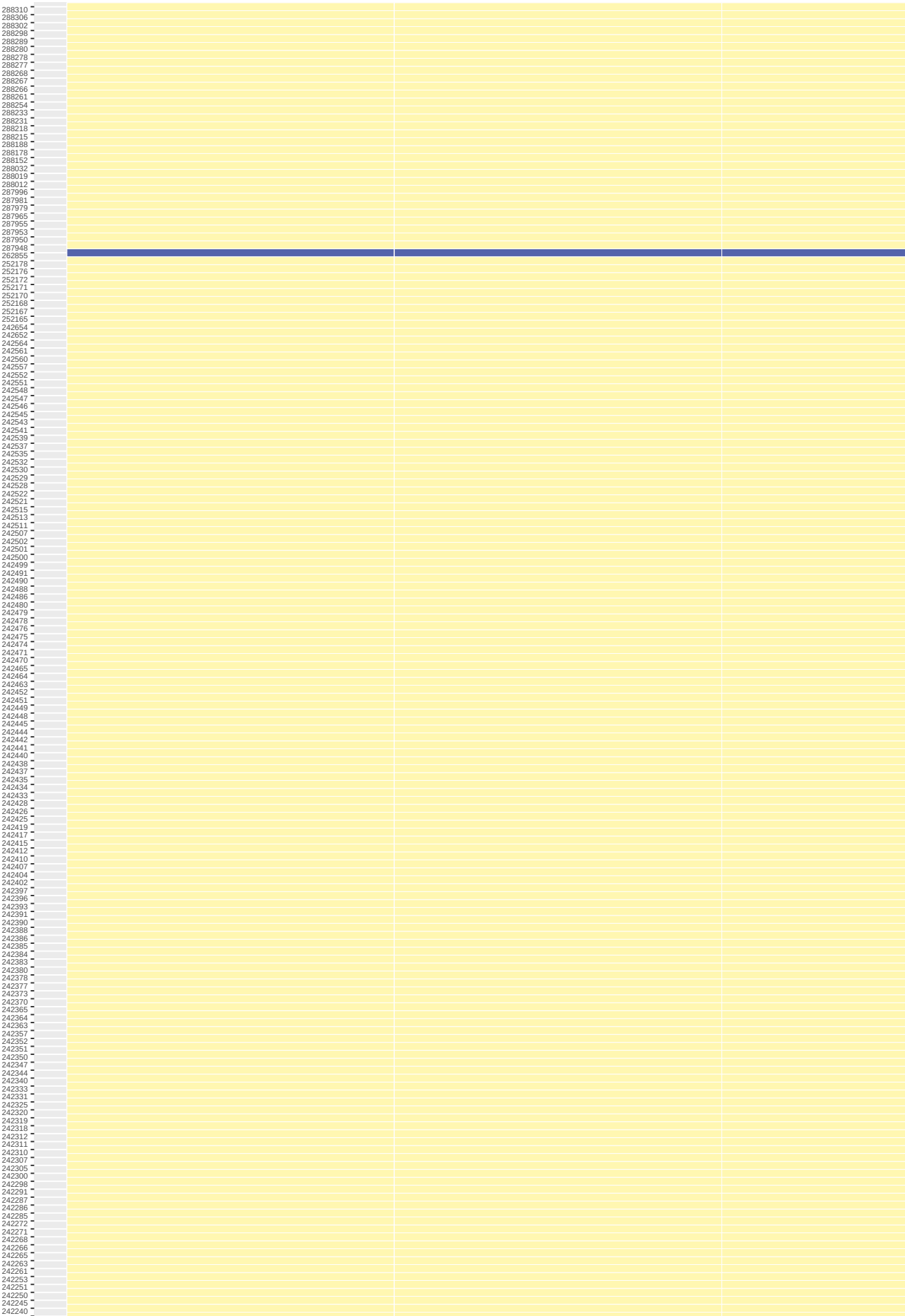
Heat plot of central pathways expansions, Needs to be phylogenetically sorted.

Here is a reference to the HeatPlot: Figure 3.2.



PPP pahtway expansions restricted to *Streptomycetaceae* family HeatPlot: Figure 3.2.

Here is a reference to the HeatPlot: Figure 3.3.



3.3 Genome Size correlations

3.3.1 Correlation between genome size and AntiSMASH products

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

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

Genome size vs Total antismash cluster coloured by order

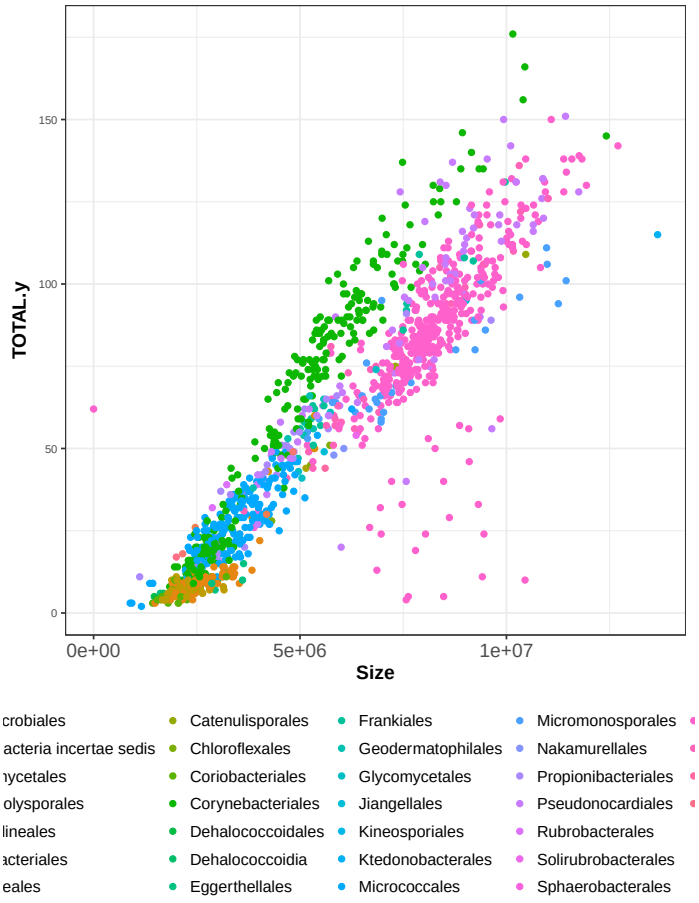


Figure 3.4: Correlation between Actinos genome size and antismash Natural products detection colored by Order

Here is a reference to Genome size vs Total antismash cluster: Figure 3.4.

Genome size vs Total antismash cluster detected splitted by order

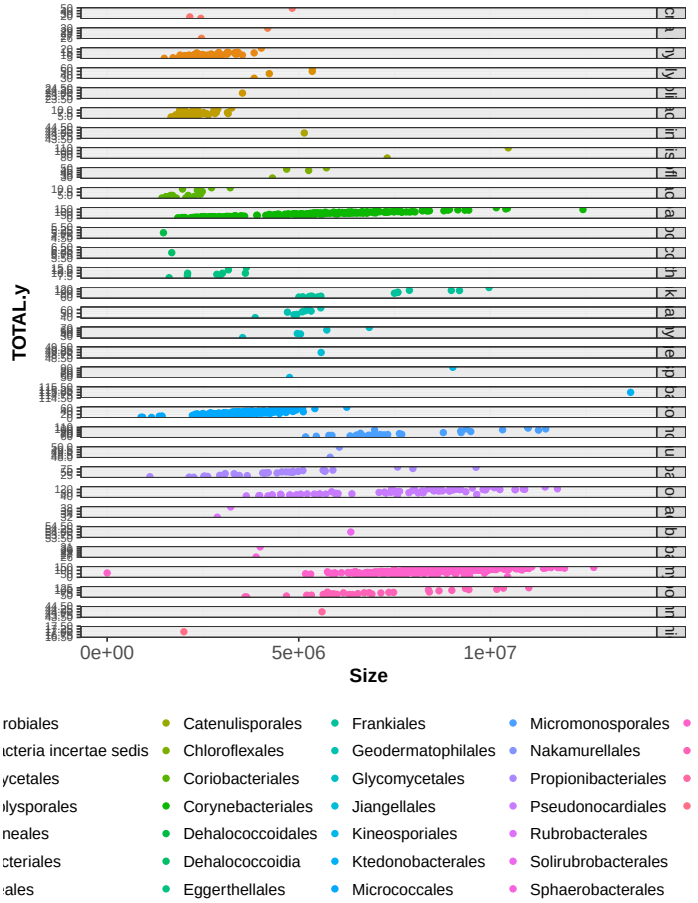


Figure 3.5: Correlation between Actinos genome size and antismash Natural products detection grided by Order

Here is a reference to Correlation between genome size and antismash Natural products detection grided by Order plot: Figure 3.5.

3.3.2 Correlation between genome size and Central pathway expansions

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

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

Genome size vs Total central pathway expansion coloured by order

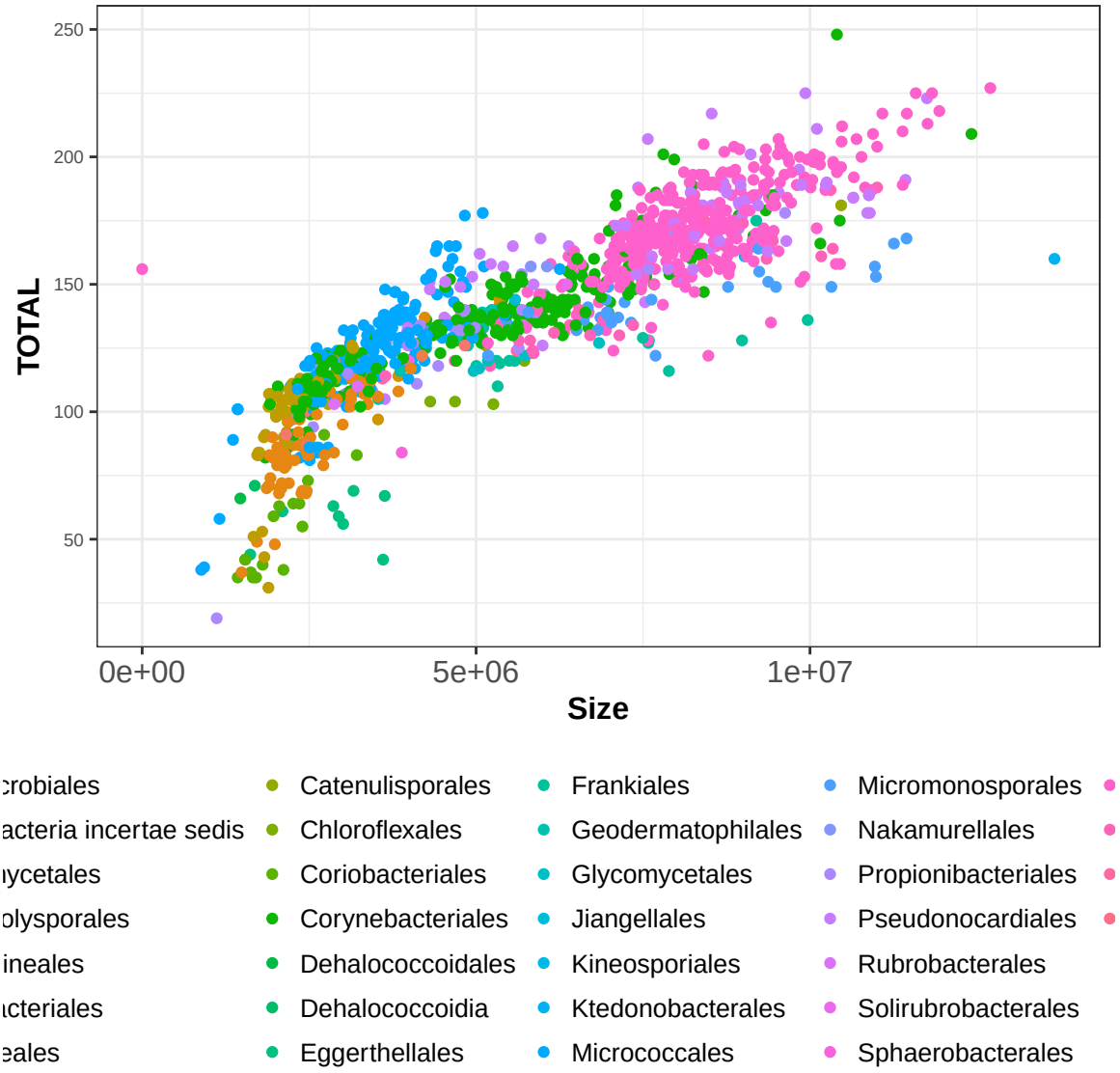


Figure 3.6: Correlation between Actinos genome size and central pathway expansions

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

Genome size vs Total central pathway expansion grided by order

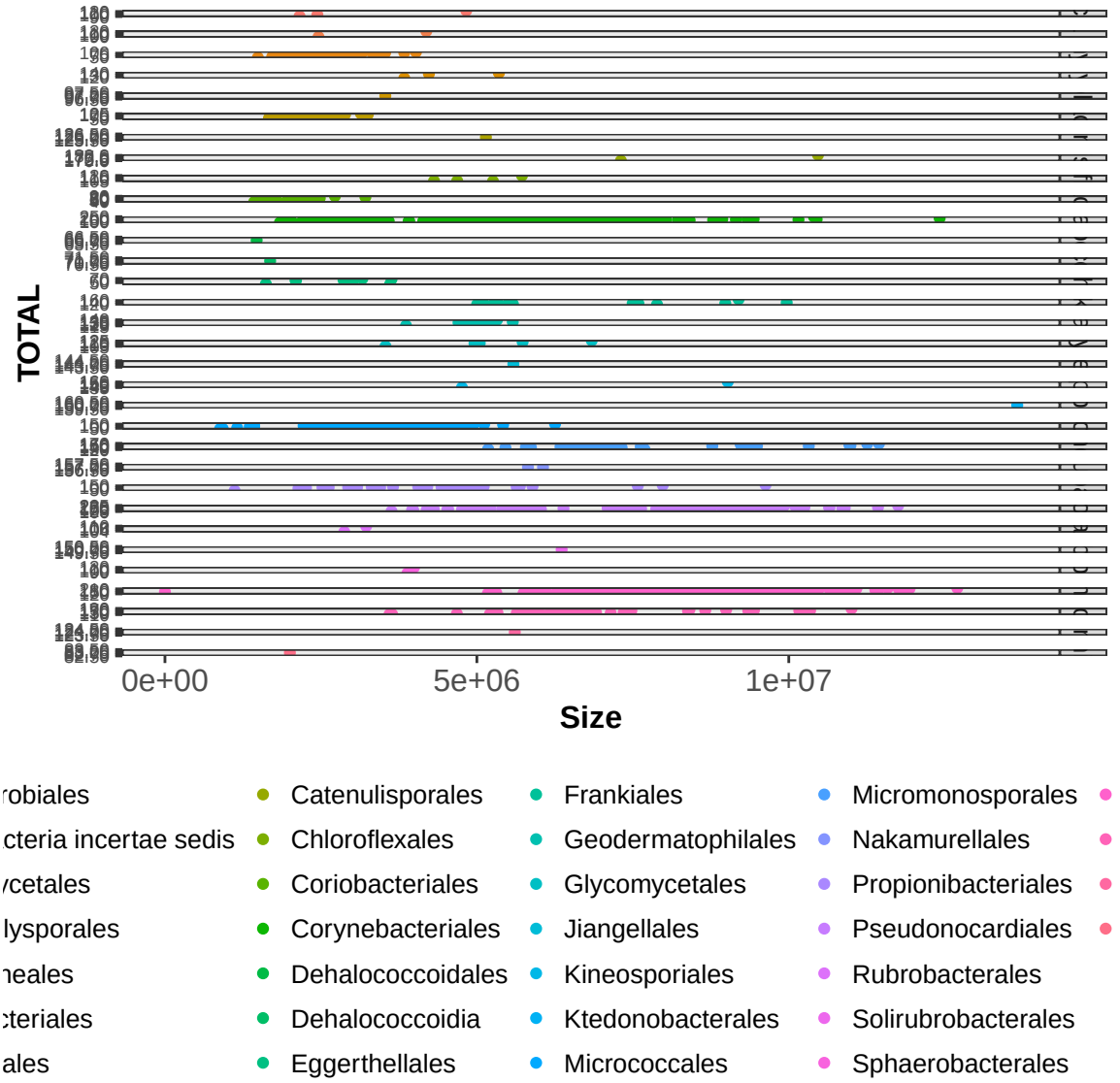


Figure 3.7: Correlation between Actinos genome size and central pathway expansions grided by order

Here is a reference to the Genome size vs Total central pathway expansion grided by order plot: Figure 3.7.

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

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

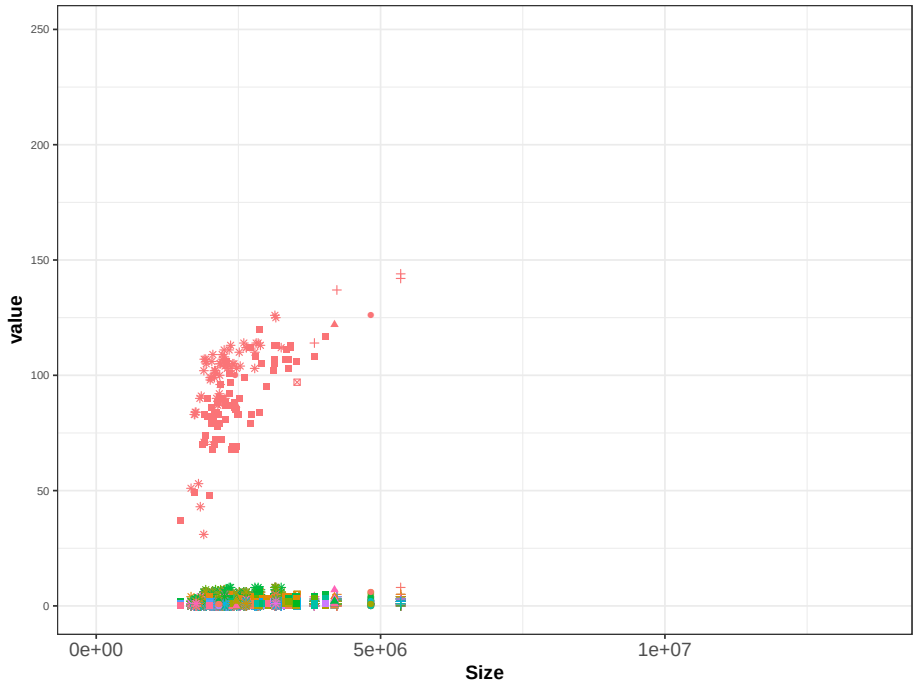


Figure 3.8: Correlation between Actinos Genome size vs Total central pathway expansion coloured by metabolic Family

Here is a reference to the Genome size vs Total central pathway expansion coloured by metabolic Family plot: Figure 3.8.

Future Work: Genome size vs Total central pathway expansion grided by metabolic Family For clarity I need to also grid and group by Metabolic Pathway

Here is a reference to Genome size vs Total central pathway expansion grided by metabolic Family plot: **??**.

3.4 Natural products

3.4.1 Natural products recruitments from EvoMining heatplot

We can see natural products recruitment after central pathways expansions colored by their kingdom.
Natural products recruited by metabolic family, colored by phylogenetic origin.

Recruitments after central pathways expansions coloured by Kingdom

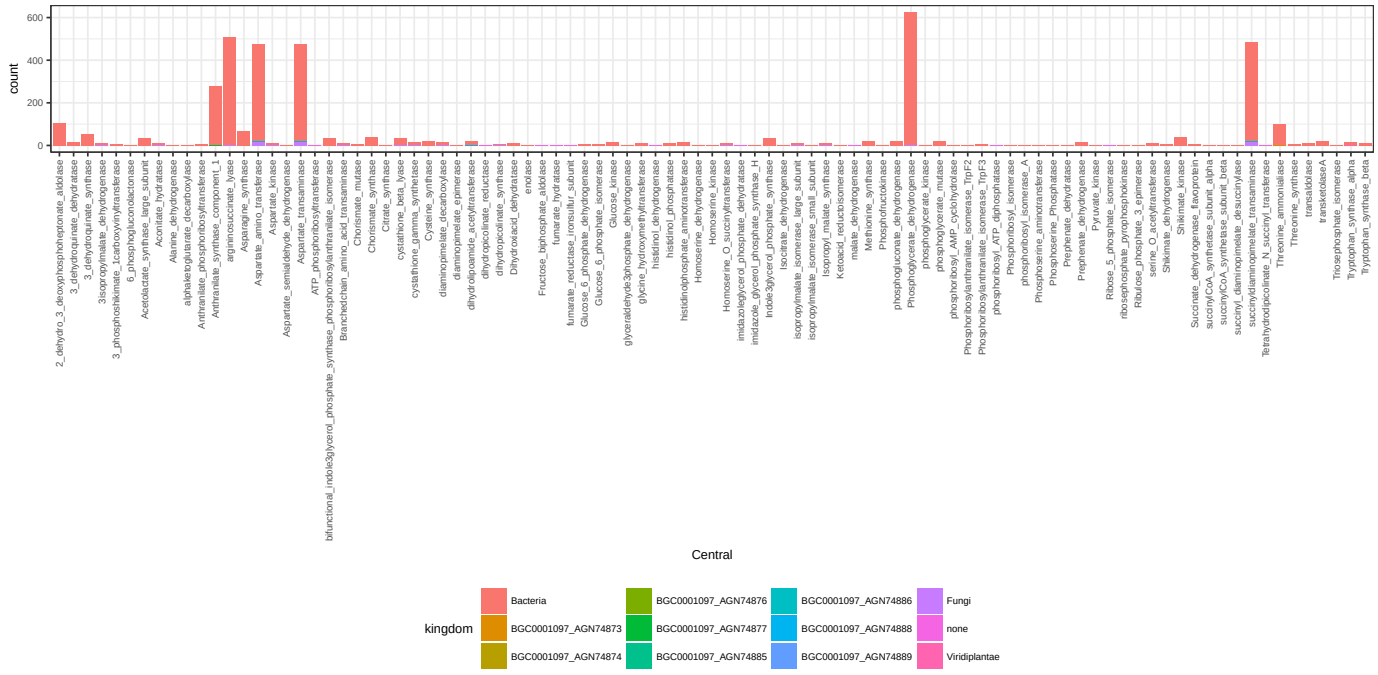


Figure 3.9: Actinos Recruitmens on central families coloured by kingdom

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

Recruitments after central pathways expansions colourd by taxonomy

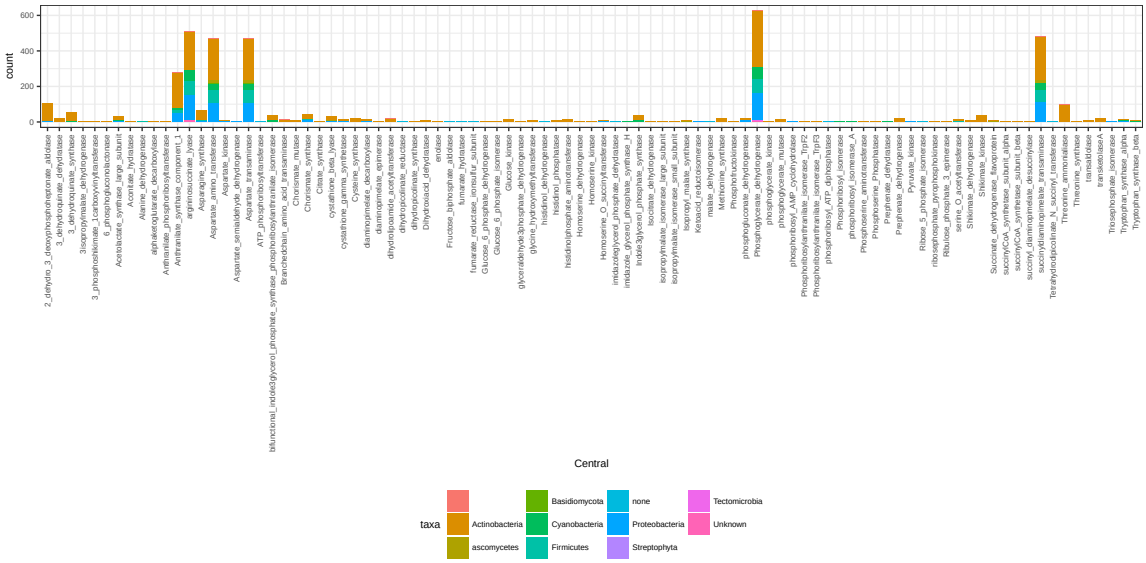


Figure 3.10: Actinos Recruitmens on central families coloured by taxonomy

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

3.5 Actinos AntiSMASH

Taxonomical diversity on Actinosbacteria Data

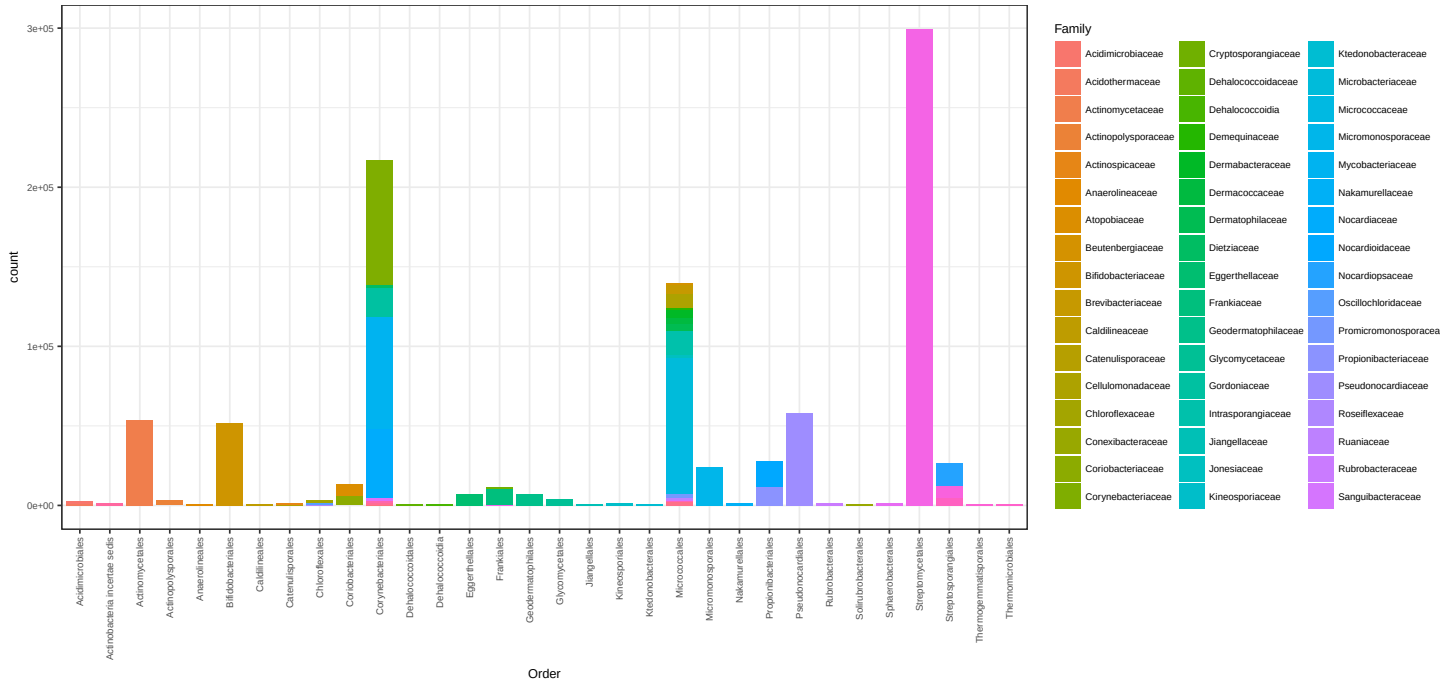


Figure 3.11: Actinos Diversity

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

Smash diversity

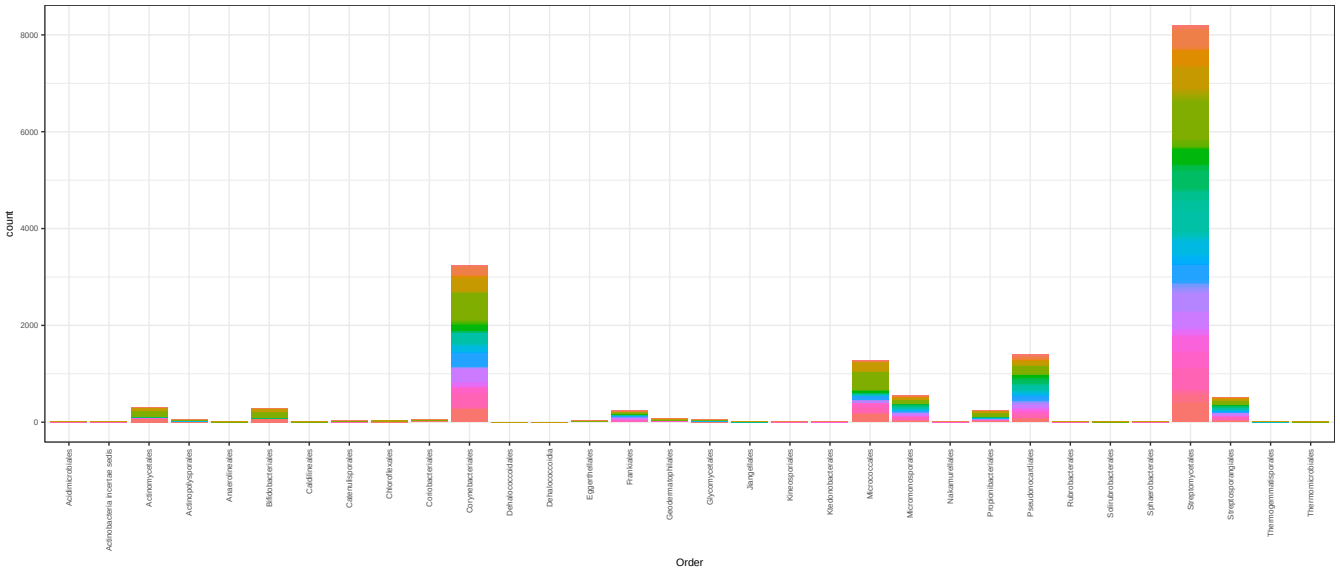


Figure 3.12: Actinos Smash Taxonomical Diversity

Here is a reference to Recruitments after central pathways expansions colourd by taxa plot: Figure 3.12.

3.5.1 AntisSMASH vs Central Expansions

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

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

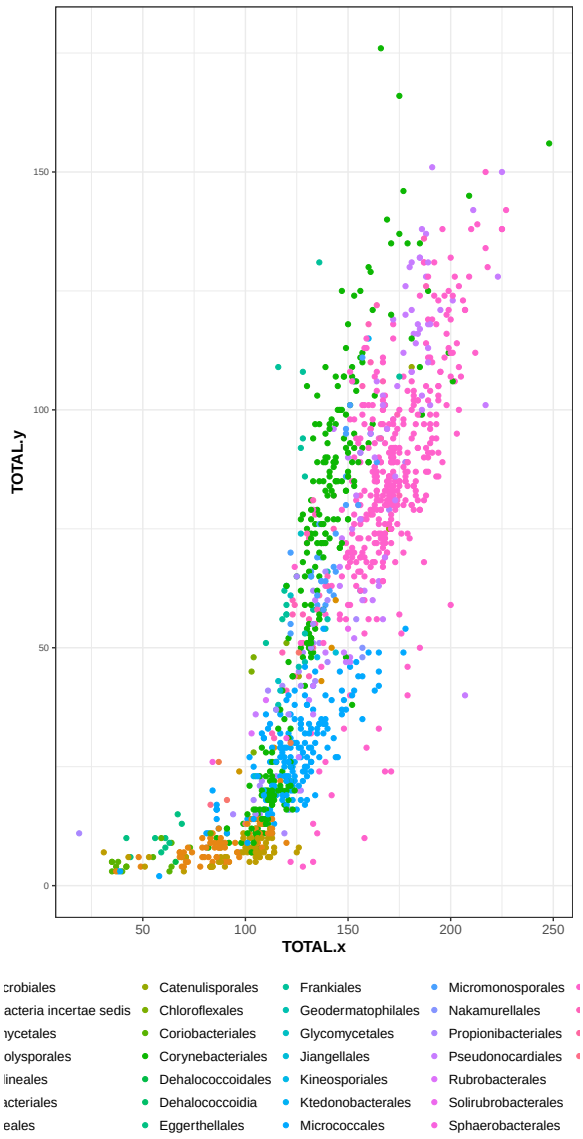


Figure 3.13: Correlation between Actinos central pathway expansions and antismash Natural products detection

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

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

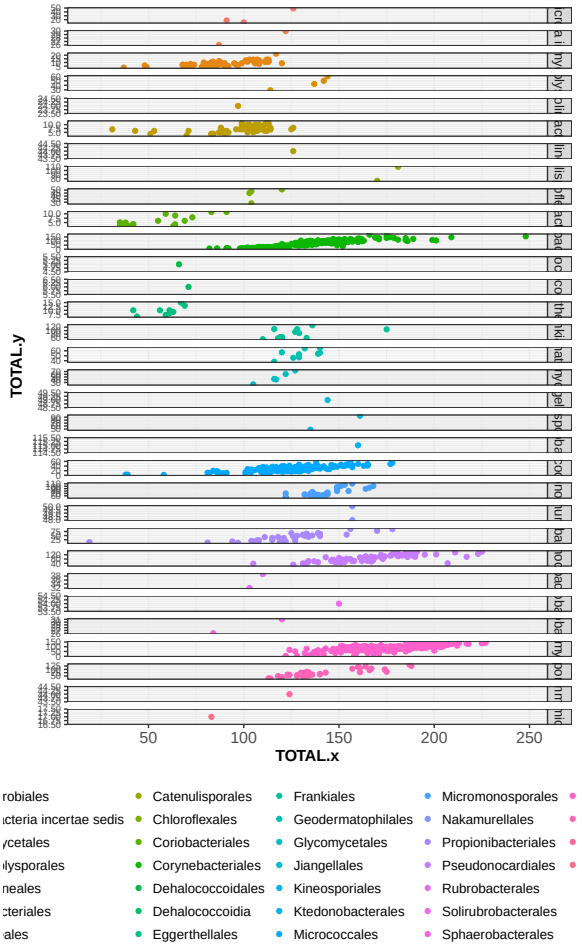


Figure 3.14: Correlation between Actinos central pathway expnasions and antismash Natural products detection

Here is a reference to the expansions vs antismash NP's clusters splitted by order plot Figure 3.14.

AntisMAsh vs Expansions by taxonomic Family

Natural products colured by family

Figure 3.15: Actinos Natural products by family

Here is a reference to the Natural products colured by family plot Figure 3.15.

3.6 Selected trees from EvoMining

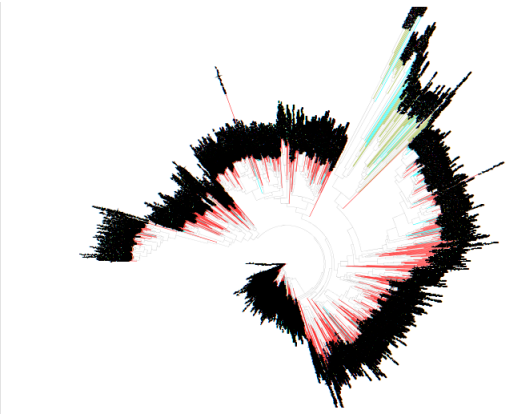


Figure 3.16: Enolase EvoMiningtree

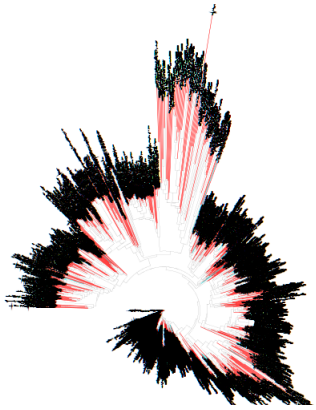


Figure 3.17: Phosphoribosyl isomerase EvoMiningtree

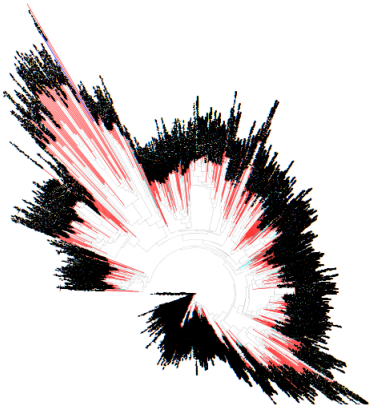


Figure 3.18: Phosphoribosyl isomerase A EvoMiningtree

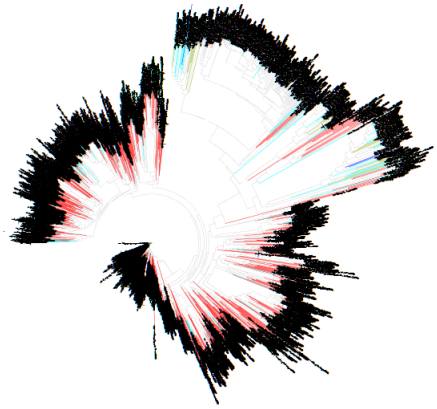


Figure 3.19: phosphoshikimate carboxyvinyltransferase EvoMiningtree

Chapter 4

Cyanobacteria EvoMining Results

Cyanobacteria phylum {Referencia}
Cyanobacteria is a photosynthetic phylum that inhabits a broad range of habitats. The broad adaptive potential is on part driven by gene-family enlargment [125] 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 [126] Example of a PriA cluster toxins[91]

Fossil record situates Cyanobacteria [126] Molecular record and metabolic propoerties at [129]

4.1 Tables

Table 4.1: Families on Cyanobacteria

Factors	Correlation between Parents & Child
GenomeDB	1245
Families	65

4.1.1 Expansions BoxPlot by metabolic family

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

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

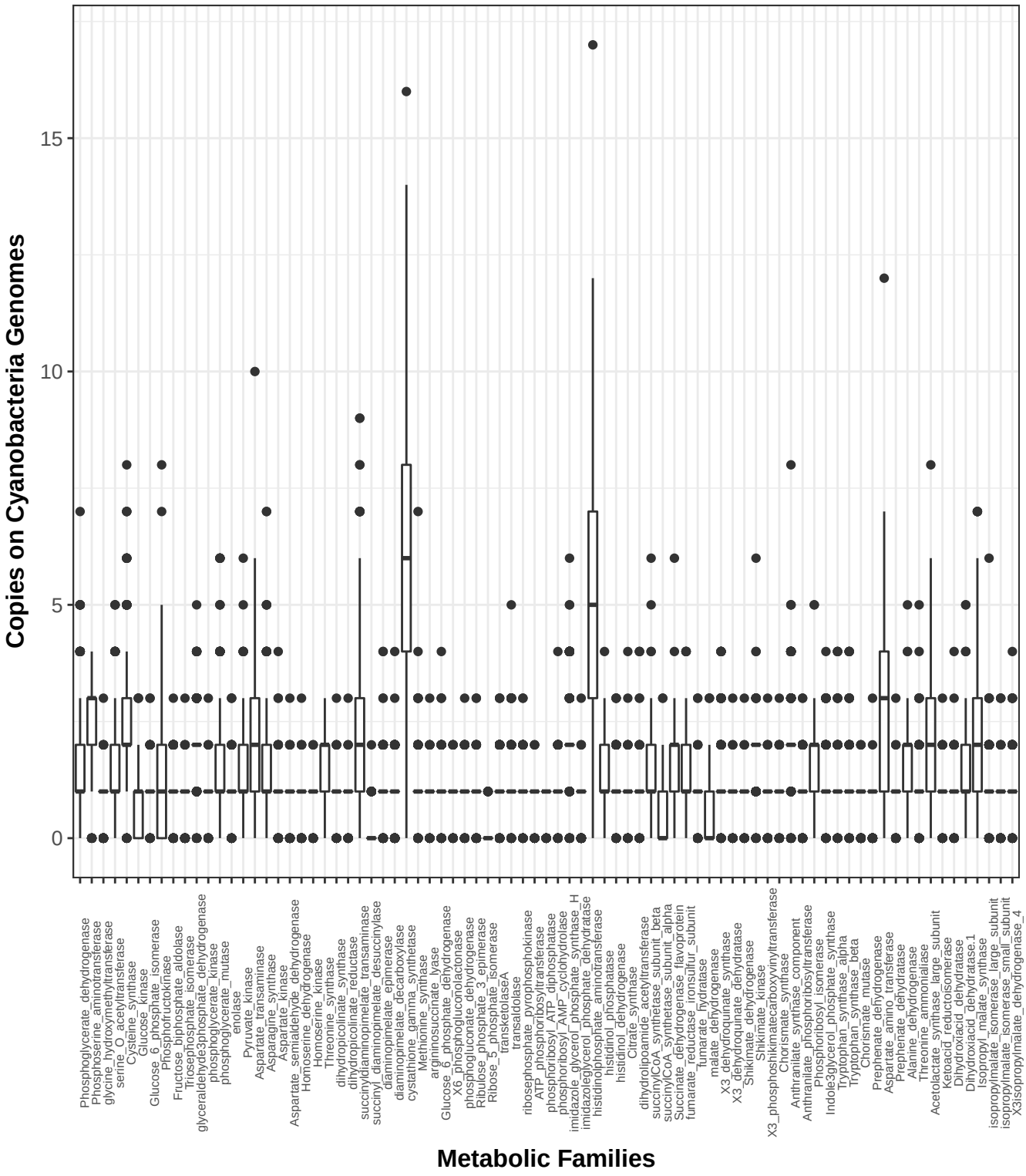


Figure 4.1: Expansions Boxplot

4.2 Central pathway expansions

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

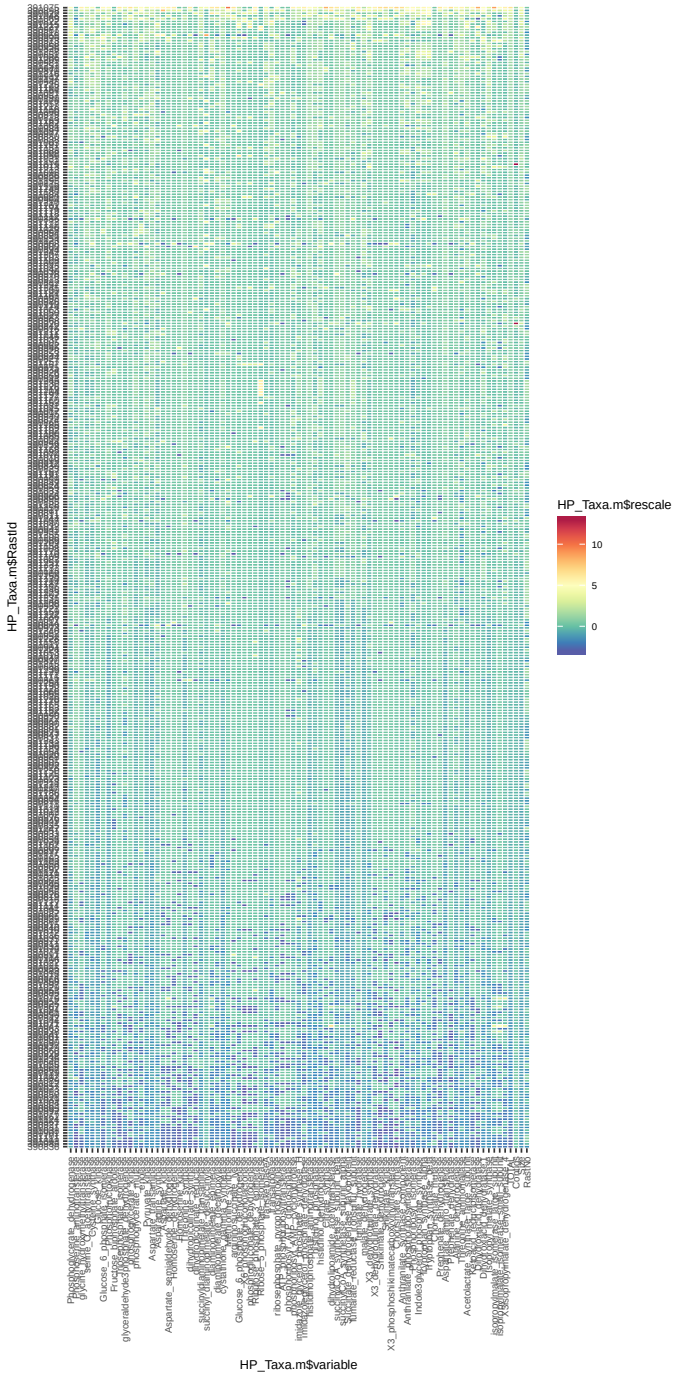


Figure 4.2: Cyanobacterial Heatplot

Here is a reference to the HeatPlot: Figure 4.2.

4.3 Genome Size correlations

4.3.1 Correlation between genome size and AntiSMASH products

Genome size vs Total antismash cluster coloured by order

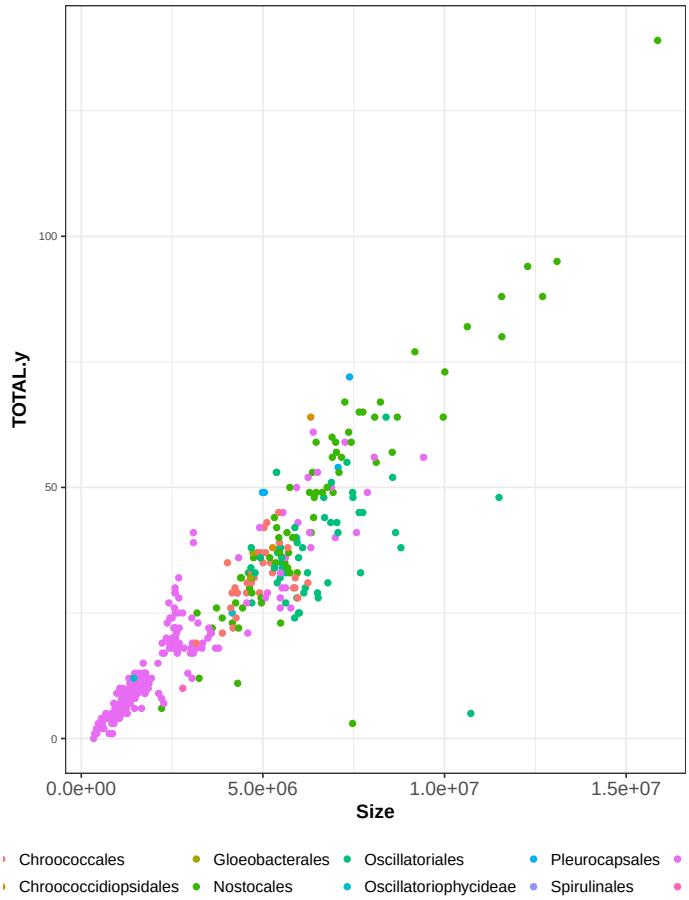


Figure 4.3: Correlation between genome size and antismash Natural products detection colored by Order

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

Genome size vs Total antismash cluster detected splitted by order

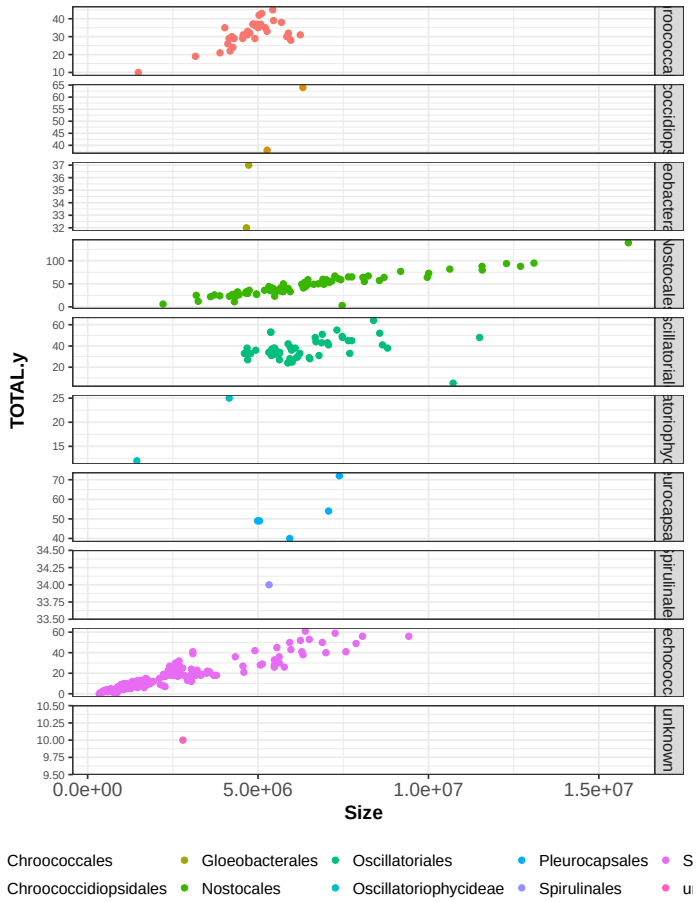


Figure 4.4: Correlation between genome size and antismash Natural products detection grided by Order

Here is a reference to Correlation between genome size and antismash Natural products detection grided by Order plot: Figure 4.4.

4.3.2 Correlation between genome size and Central pathway expansions

Genome size vs Total central pathway expansion coloured by order

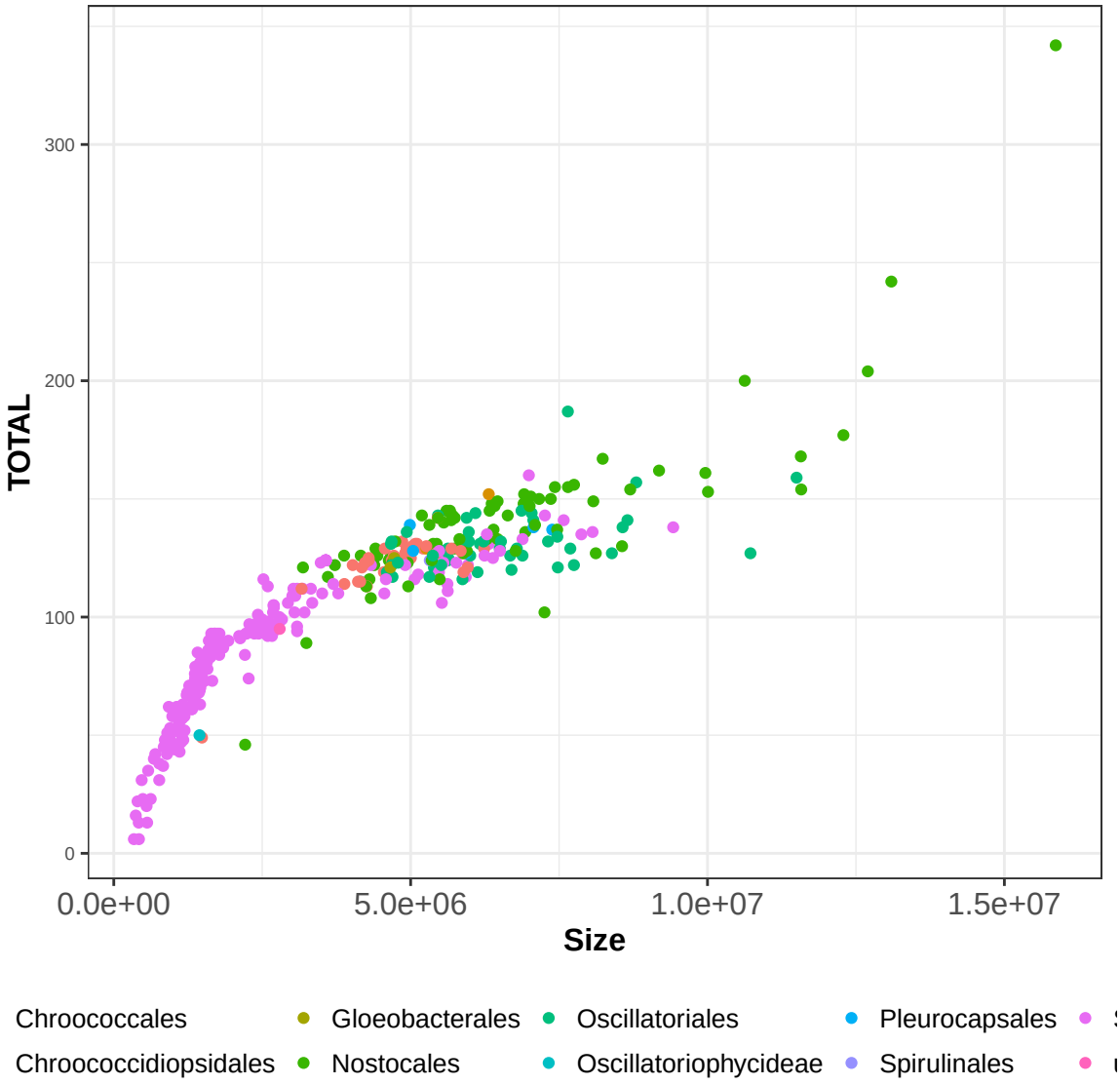


Figure 4.5: Correlation between genome size and central pathway expansions

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

Genome size vs Total central pathway expansion grided by order

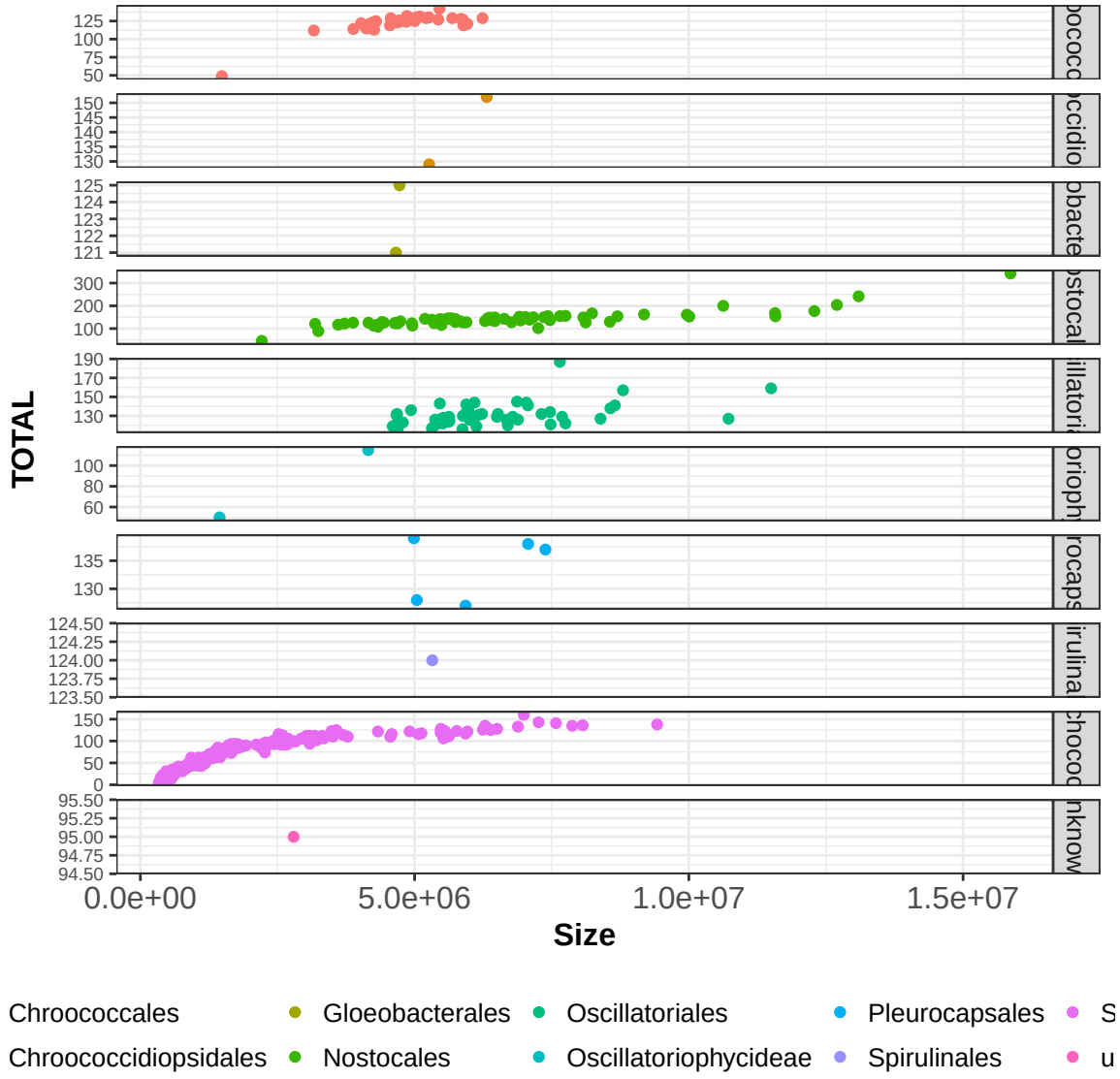


Figure 4.6: Correlation between genome size and central pathway expansions grided by order

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

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

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

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

Genome size vs Total central pathway expansion coloured by metabolic Family

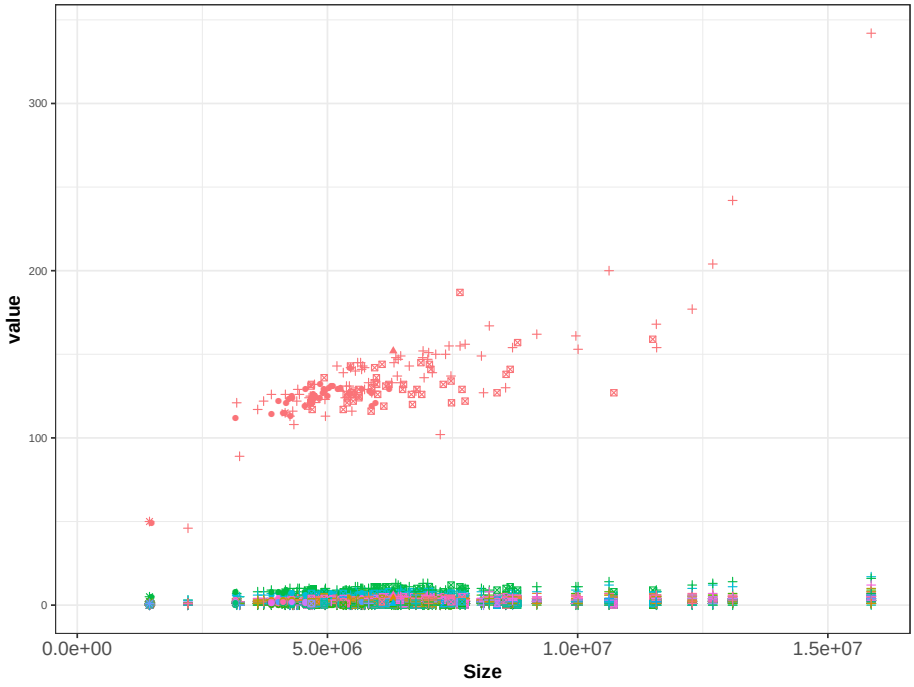


Figure 4.7: Correlation between Genome size vs Total central pathway expansion coloured by metabolic Family

Here is a reference to the Genome size vs Total central pathway expansion coloured by metabolic Family plot: Figure 4.7.

Future Work: Genome size vs Total central pathway expansion grided by metabolic Family For clarity I need to also grid and group by Metabolic Pathway

Here is a reference to Genome size vs Total central pathway expansion grided by metabolic Family plot: **??**.

4.4 Natural products

4.4.1 Natural products recruitments from EvoMining heatplot

We can see natural products recruitment after central pathways expansions colored by their kingdom.
Natural products recruited by metabolic family, colored by phylogenetic origin.

Recruitments after central pathways expansions coloured by Kingdom

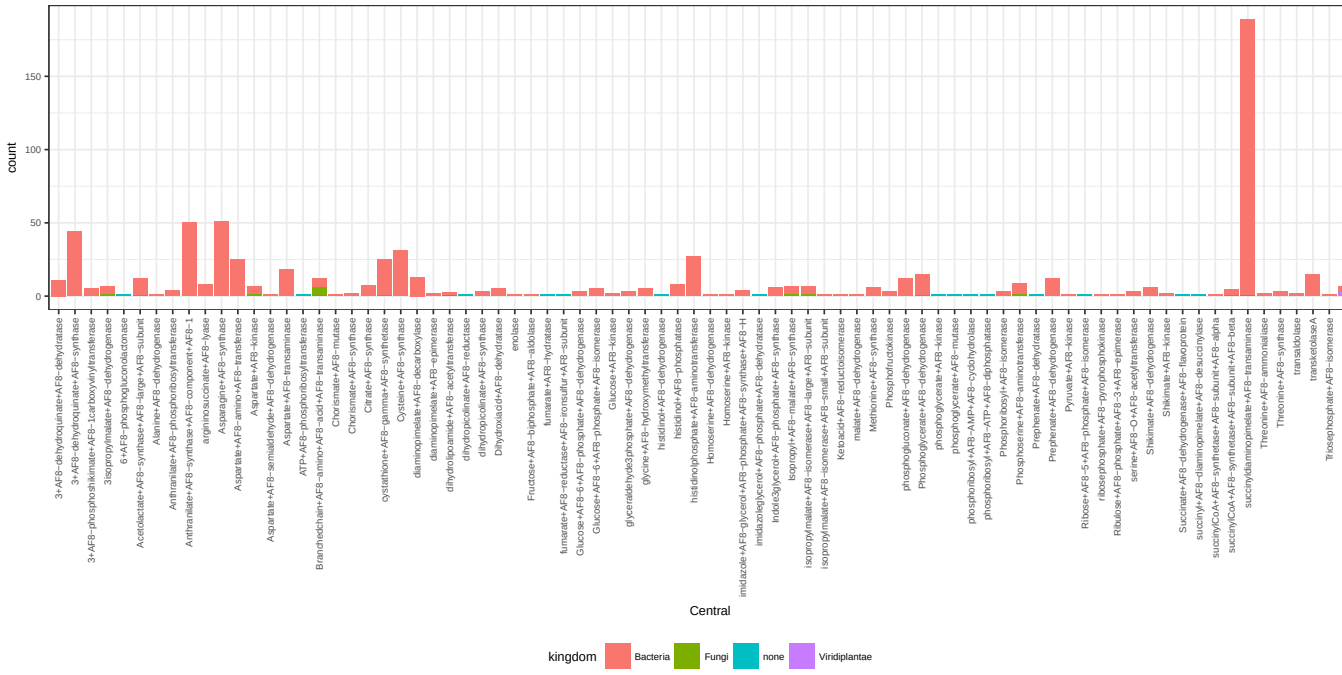


Figure 4.8: Recruitmens on central families coloured by kingdom

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

Recruitments after central pathways expansions colourd by taxonomy

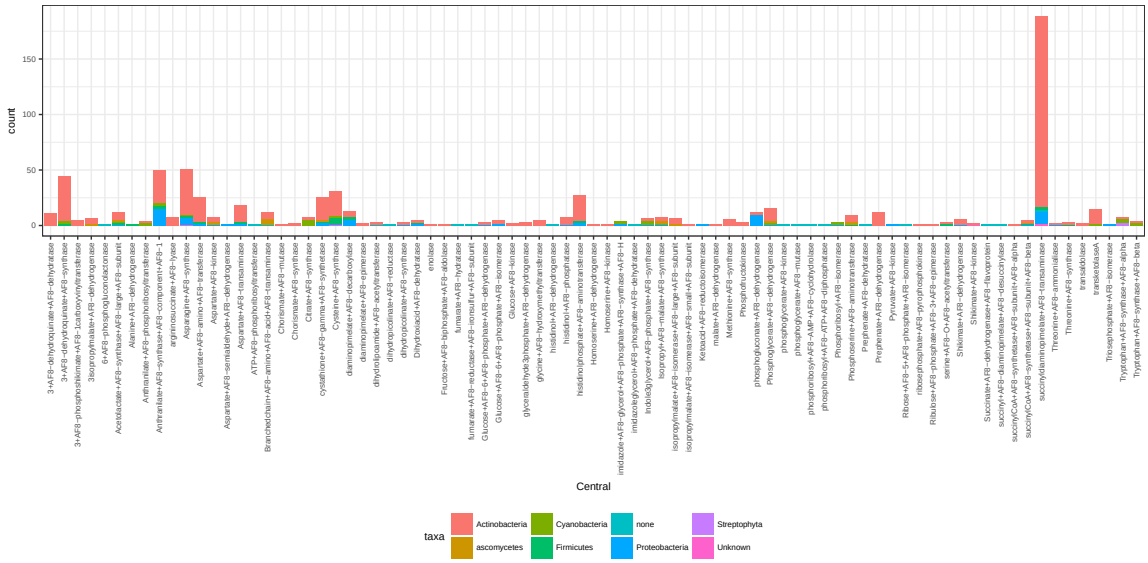


Figure 4.9: Recruitmens on central families coloured by taxonomy

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

4.5 Cyanobacterias AntiSMASH

Taxonomical diversity on Cyanobacteria Data

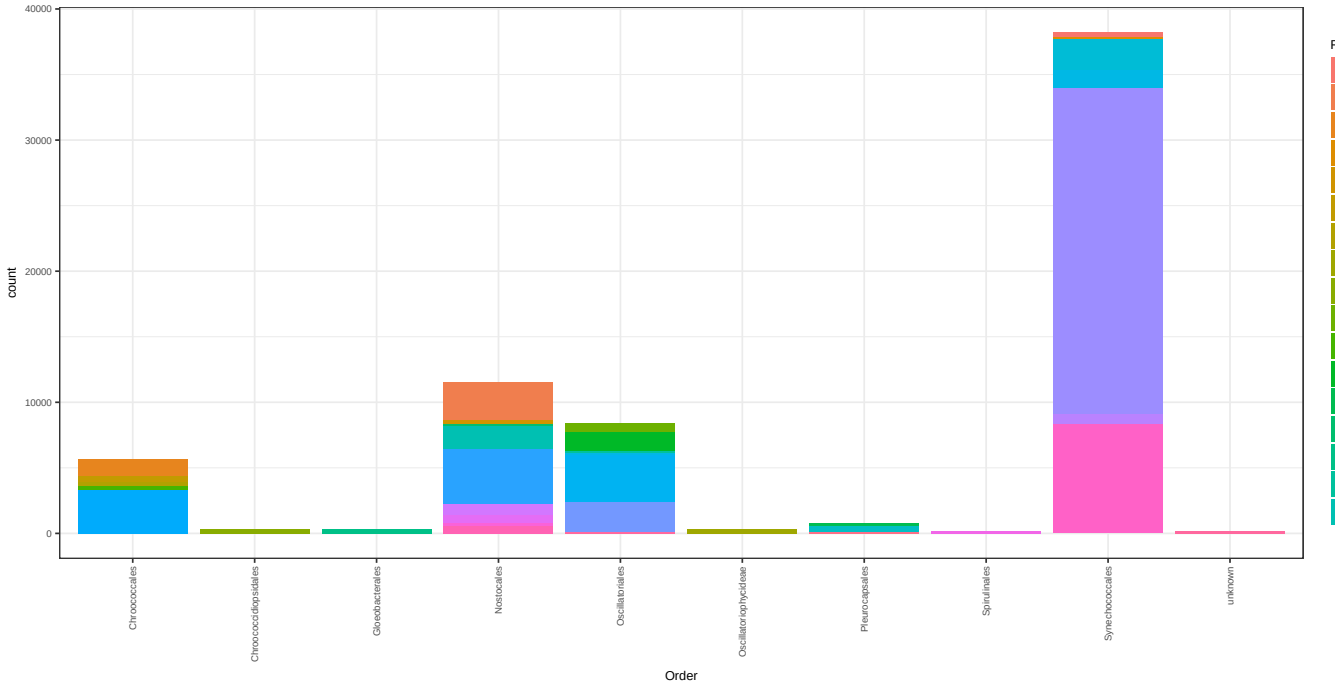


Figure 4.10: Diversity

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

Smash diversity

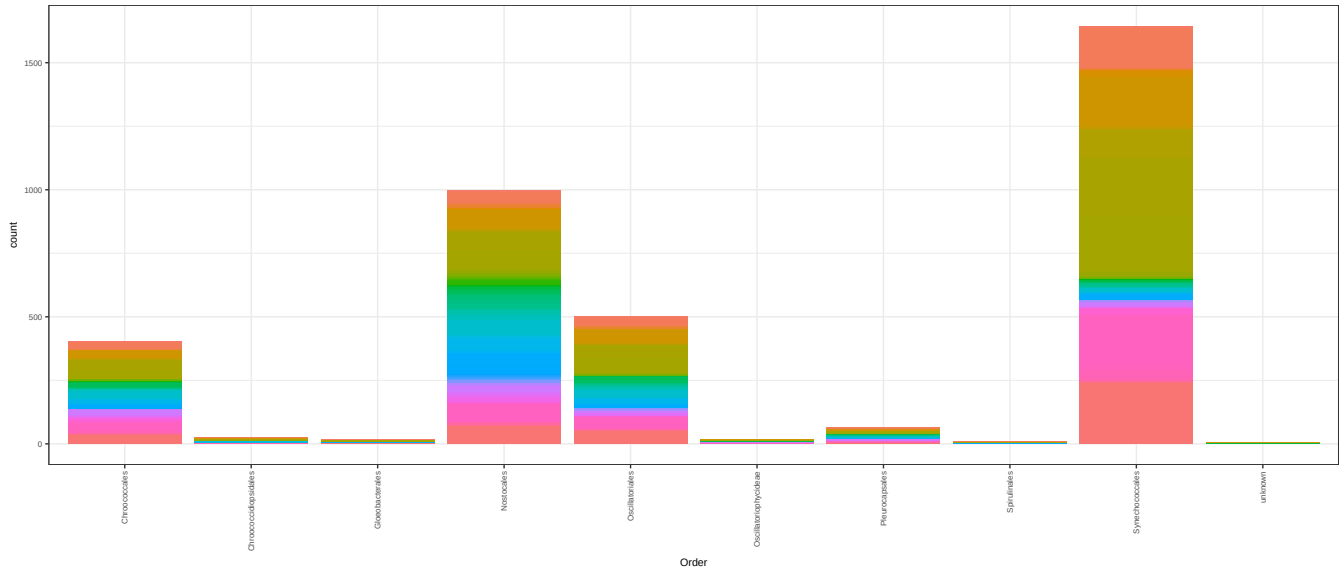


Figure 4.11: Smash

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

4.5.1 AntisSMASH vs Central Expansions

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

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

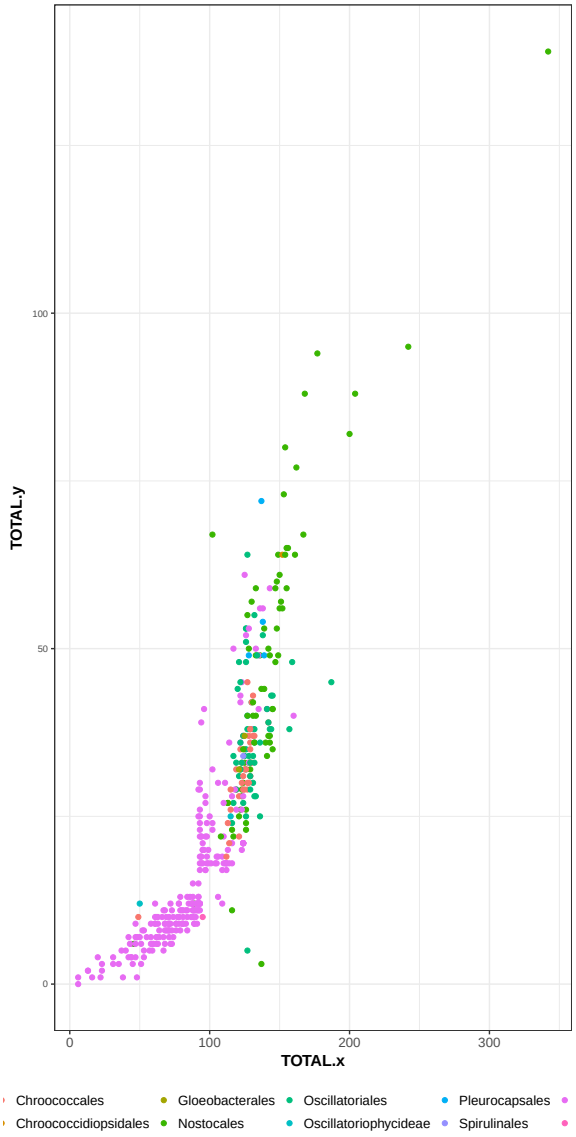


Figure 4.12: Correlation between central pathway axpnasions and antismash Natural products detection

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

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

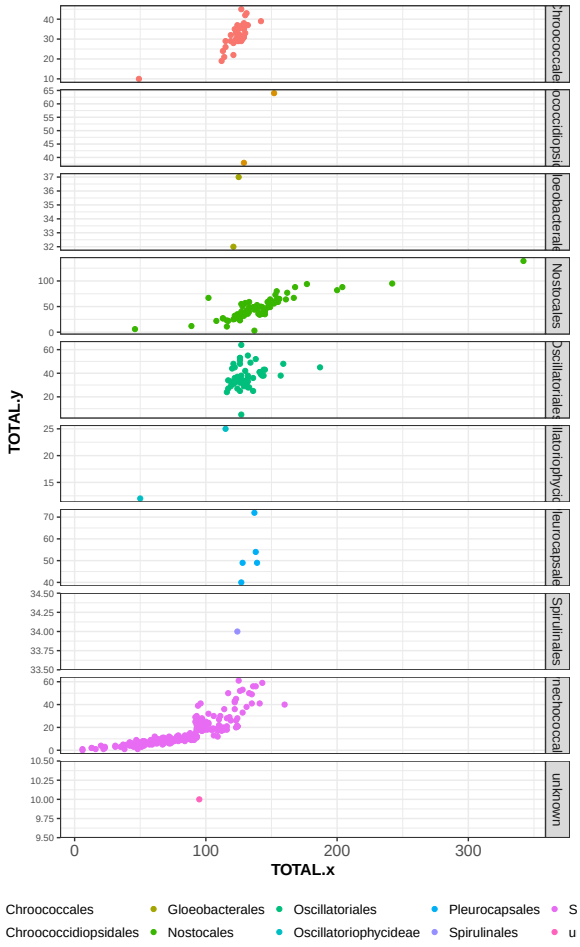


Figure 4.13: Correlation between central pathway axpnasions and antismash Natural products detection

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

AntisMAsh vs Expansions by taxonomic Family

Natural products colured by family

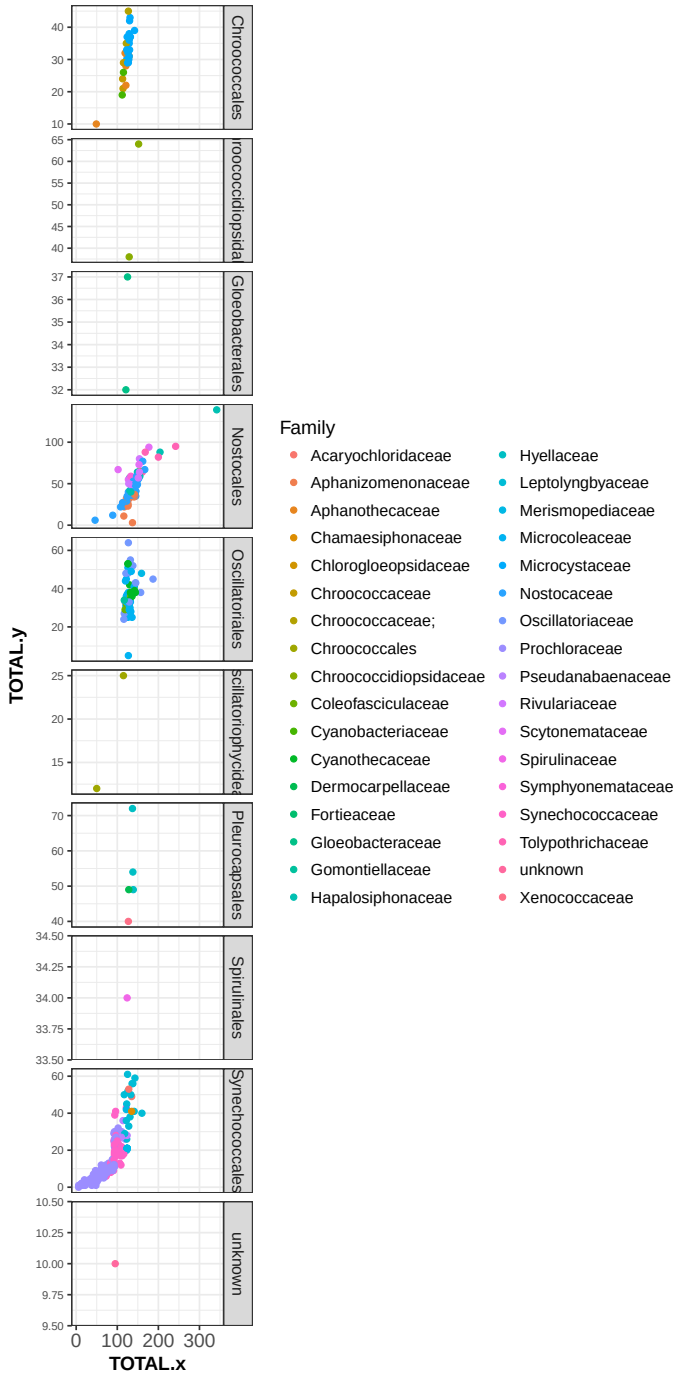


Figure 4.14: Natural products by family

Here is a reference to the Natural products colured by family plot Figure 4.14.

4.6 Selected trees from EvoMining

Phosphoribosyl_isomerase_3 family
Figure from EvoMining

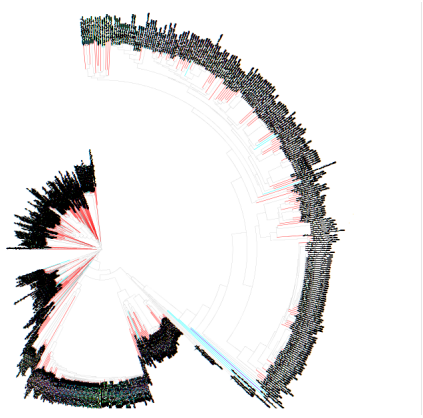


Figure 4.15: Phosphoribosyl isomerase EvoMiningtree

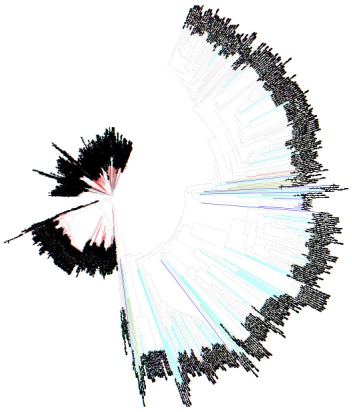


Figure 4.16: Phosphoglycerate dehydrogenase EvoMiningtree

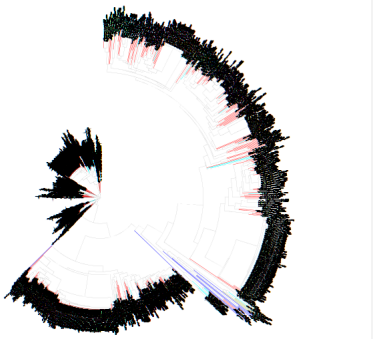


Figure 4.17: Phosphoserine aminotransferase EvoMiningtree

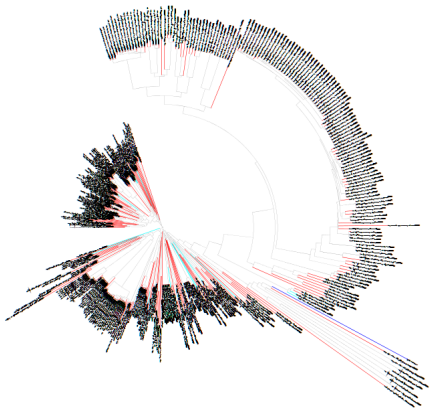


Figure 4.18: Triosephosphate isomerase EvoMiningtree

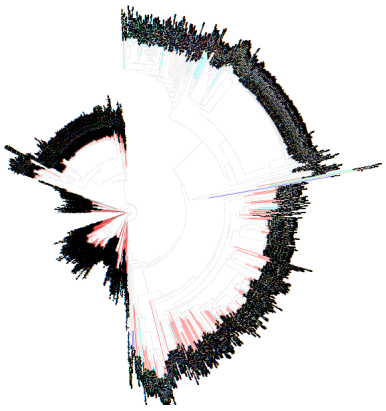


Figure 4.19: glyceraldehyde3phosphate dehydrogenase EvoMiningtree

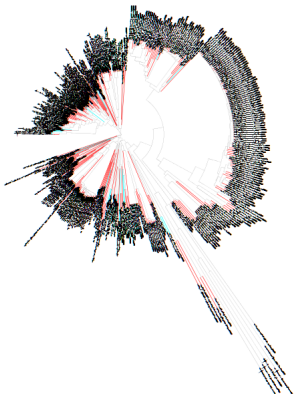


Figure 4.20: phosphoglycerate kinase EvoMiningtree

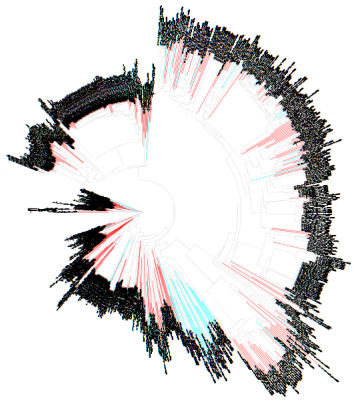


Figure 4.21: phosphoglycerate mutaseEvoMiningtree

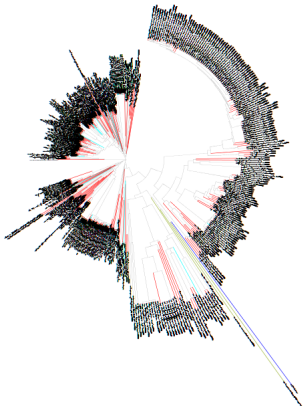


Figure 4.22: enolase EvoMiningtree

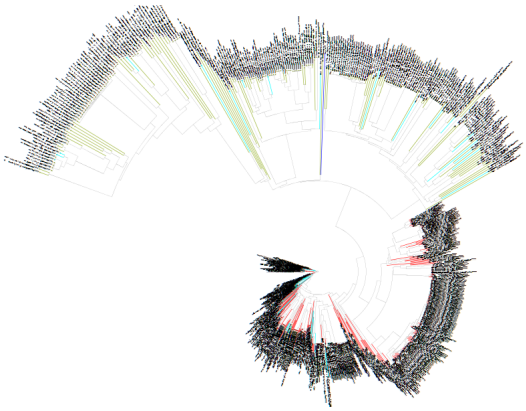


Figure 4.23: Pyruvate kinase EvoMiningtree

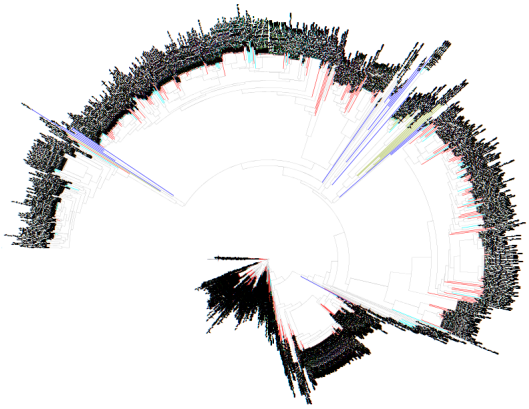


Figure 4.24: Aspartate transaminase EvoMiningtree

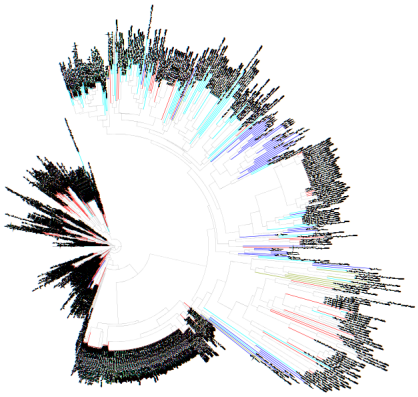


Figure 4.25: Asparagine synthase EvoMiningtree

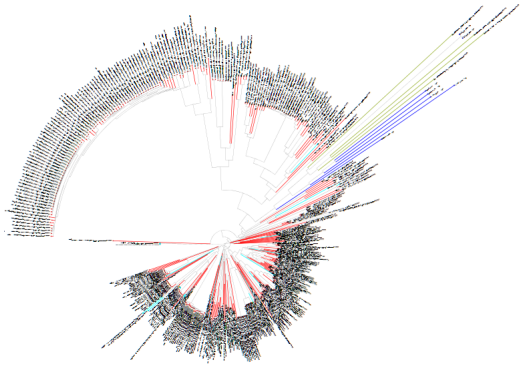


Figure 4.26: Aspartate kinase EvoMiningtree

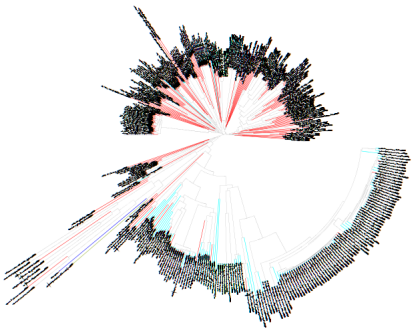


Figure 4.27: Aspartate semialdehyde dehydrogenase EvoMiningtree

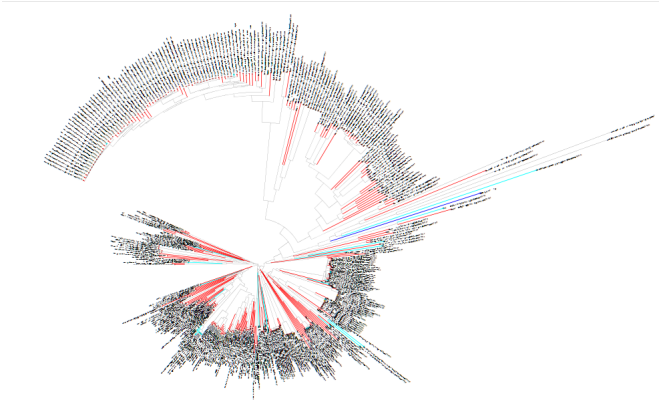


Figure 4.28: Homoserine dehydrogenase EvoMiningtree

Conclusion

Idea de Rosario -ver dell cluster de saxitoxin cuantos pasos se necesitron para llegar ahi.
-A donde se iria el resultado de abrir el GMP
-Otra vez, que Actinos tienen FolE

If we don't want Conclusion to have a chapter number next to it, we can add the `{.unnumbered}` attribute. This has an unintended consequence of the sections being labeled as 3.6 for example though instead of 4.1. The LaTeX 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 readibility 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 Rmakdown 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/`

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 expanssions

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. doi:10.1146/annurev-biochem-030409-143718

2. Copley SD. Enzymes with extra talents: Moonlighting functions and catalytic promiscuity. Current Opinion in Chemical Biology. 2003;7: 265–272. doi:10.1016/S1367-5931(03)00032-2

3. Hult K, Berglund P. Enzyme promiscuity: Mechanism and applications. Trends in Biotechnology. 2007;25: 231–238. doi:10.1016/j.tibtech.2007.03.002

4. O'Brien PJ, Herschlag D. Catalytic promiscuity and the evolution of new enzymatic activities. Chemistry & Biology. 1999;6: R91–R105. doi:10.1016/S1074-5521(99)80033-7

5. Barona Gómez F, Hodgson DA. Occurrence of a putative ancient like isomerase involved in histidine and tryptophan biosynthesis. EMBO reports. 2003;4: 296–300. doi:10.1038/sj.embor.embor771

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. 2014;82: 887–896. doi:10.1002/prot.24575

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. 2007;21: 368–372. doi: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. 2004;59: 5051–5060. doi:10.1016/j.ces.2004.09.021

9. Glasner ME, Gerlt JA, Babbitt PC. Evolution of enzyme superfamilies. Current Opinion in Chemical Biology. 2006;10: 492–497. doi:10.1016/j.cbpa.2006.08.012

10. Baier F, Copp JN, Tokuriki N. Evolution of Enzyme Superfamilies: Comprehensive Exploration of Sequence–Function Relationships. Biochemistry. 2016;55: 6375–

6388. doi:10.1021/acs.biochem.6b00723

11. Bloom JD, Romero PA, Lu Z, Arnold FH. Neutral genetic drift can alter promiscuous protein functions, potentially aiding functional evolution. Biology Direct. 2007;2: 17. doi:10.1186/1745-6150-2-17

12. Nath A, Atkins WM. A Quantitative Index of Substrate Promiscuity. Biochemistry. 2008;47: 157–166. doi:10.1021/bi701448p

13. Zou T, Risso VA, Gavira JA, Sanchez-Ruiz JM, Ozkan SB. Evolution of Conformational Dynamics Determines the Conversion of a Promiscuous Generalist into a Specialist Enzyme. Molecular Biology and Evolution. 2015;32: 132–143. doi:10.1093/molbev/msu281

14. Firn RD, Jones CG. A Darwinian view of metabolism: Molecular properties determine fitness. Journal of Experimental Botany. 2009;60: 719–726. doi:10.1093/jxb/erp002

15. Jia B, Cheong G-W, Zhang S. Multifunctional enzymes in archaea: Promiscuity and moonlight. Extremophiles. 2013;17: 193–203. doi:10.1007/s00792-012-0509-1

16. Aharoni A, Gaidukov L, Khersonsky O, Gould SM, Roodveldt C, Tawfik DS. The 'evolvability' of promiscuous protein functions. Nature Genetics. 2005;37: 73–76. doi:10.1038/ng1482

17. Jensen. Enzyme Recruitment in Evolution of New Function. Annual Review of Microbiology. 1976;30: 409–425. doi: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. 2014;289: 30229–30236. doi:10.1074/jbc.R114.572990

19. Dean AM, Thornton JW. Mechanistic approaches to the study of evolution. Nature reviews Genetics. 2007;8: 675–688. doi:10.1038/nrg2160

20. Nobeli I, Favia AD, Thornton JM. Protein promiscuity and its implications for biotechnology. Nature Biotechnology. 2009;27: 157–167. doi:10.1038/nbt1519

21. Hopkins AL. Drug discovery: Predicting promiscuity. Nature. 2009;462: 167–168. doi:10.1038/462167a

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. 2010;38: 2195–2203. doi:10.1124/dmd.110.034645

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. 2011;39: D1060–D1066. doi:10.1093/nar/gkq1037

24. Zhang W, Dourado DFAR, Fernandes PA, Ramos MJ, Mannervik B. Multidimensional epistasis and fitness landscapes in enzyme evolution. Biochemical Journal. 2012;445: 39–46. doi:10.1042/BJ20120136

25. Sanchez-Ruiz JM. On promiscuity, changing environments and the possibility of replaying the molecular tape of life. Biochemical Journal. 2012;445: e1–e3. doi:10.1042/BJ20120806

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. 2015;83: 1625–1631. doi:10.1002/prot.24847

27. Patrick WM, Quandt EM, Swartzlander DB, Matsumura I. Multicopy Suppression Underpins Metabolic Evolvability. Molecular Biology and Evolution. 2007;24: 2716–2722. doi:10.1093/molbev/msm204

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. 2014;111: 11762–11767. doi:10.1073/pnas.1406102111

29. Linster CL, Van Schaftingen E, Hanson AD. Metabolite damage and its repair or pre-emption. Nature Chemical Biology. 2013;9: 72–80. doi:10.1038/nchembio.1141

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. 2015;32: 100–108. doi:10.1093/molbev/msu271

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. 2013;20: 796–805. doi:10.1016/j.chembiol.2013.04.013

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. 2014;5. doi:10.1128/mBio.01378-14

33. Soskine M, Tawfik DS. Mutational effects and the evolution of new protein functions. Nature Reviews Genetics. 2010;11: 572–582. doi:10.1038/nrg2808

34. Halachev MR, Loman NJ, Pallen MJ. Calculating Orthologs in Bacteria and Archaea: A Divide and Conquer Approach. PLOS ONE. 2011;6: e28388.

doi:10.1371/journal.pone.0028388

35. Kislyuk AO, Haegeman B, Bergman NH, Weitz JS. Genomic fluidity: An integrative view of gene diversity within microbial populations. BMC Genomics. 2011;12: 32. doi:10.1186/1471-2164-12-32

36. Pearson H. Prehistoric proteins: Raising the dead. Nature News. 2012;483: 390. doi:10.1038/483390a

37. Hughes AL. The Evolution of Functionally Novel Proteins after Gene Duplication. Proceedings of the Royal Society of London B: Biological Sciences. 1994;256: 119–124. doi:10.1098/rspb.1994.0058

38. Treangen TJ, Rocha EPC. Horizontal Transfer, Not Duplication, Drives the Expansion of Protein Families in Prokaryotes. PLOS Genetics. 2011;7: e1001284. doi:10.1371/journal.pgen.1001284

39. Overbeek R, Fonstein M, D'Souza M, Pusch GD, Maltsev N. The use of gene clusters to infer functional coupling. Proceedings of the National Academy of Sciences. 1999;96: 2896–2901. doi:10.1073/pnas.96.6.2896

40. Zhao S, Sakai A, Zhang X, Vetting MW, Kumar R, Hillerich B, et al. Prediction and characterization of enzymatic activities guided by sequence similarity and genome neighborhood networks. eLife. 2014;3: e03275. doi:10.7554/eLife.03275

41. Zhao S, Kumar R, Sakai A, Vetting MW, Wood BM, Brown S, et al. Discovery of new enzymes and metabolic pathways by using structure and genome context. Nature. 2013;502: 698–702. doi:10.1038/nature12576

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. 2015;8: 239–252. doi:10.1111/1751-7915.12175

43. Szklarczyk D, Franceschini A, Wyder S, Forslund K, Heller D, Huerta-Cepas J, et al. STRING v10: Protein–protein interaction networks, integrated over the tree of life. Nucleic Acids Research. 2015;43: D447–D452. doi:10.1093/nar/gku1003

44. Snel B, Lehmann G, Bork P, Huynen MA. STRING: A web-server to retrieve and display the repeatedly occurring neighbourhood of a gene. Nucleic Acids Research. 2000;28: 3442–3444. Available: `http://www.ncbi.nlm.nih.gov/pmc/articles/PMC110752/`

45. Aziz RK, Bartels D, Best AA, DeJongh M, Disz T, Edwards RA, et al. The RAST Server: Rapid Annotations using Subsystems Technology. BMC Genomics. 2008;9: 75. doi:10.1186/1471-2164-9-75

46. Overbeek R, Olson R, Pusch GD, Olsen GJ, Davis JJ, Disz T, et al. The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST).

Nucleic Acids Research. 2014;42: D206–D214. doi:10.1093/nar/gkt1226

47. Medema MH, Fischbach MA. Computational approaches to natural product discovery. Nature Chemical Biology. 2015;11: 639–648. doi:10.1038/nchembio.1884

48. Noda-García L, Camacho-Zarco AR, Medina-Ruíz S, Gaytán P, Carrillo-Tripp M, Fülöp V, et al. Evolution of Substrate Specificity in a Recipient's Enzyme Following Horizontal Gene Transfer. Molecular Biology and Evolution. 2013;30: 2024–2034. doi:10.1093/molbev/mst115

49. Carbonell P, Faulon J-L. Molecular signatures-based prediction of enzyme promiscuity. Bioinformatics. 2010;26: 2012–2019. doi:10.1093/bioinformatics/btq317

50. Cheng X-Y, Huang W-J, Hu S-C, Zhang H-L, Wang H, Zhang J-X, et al. A Global Characterization and Identification of Multifunctional Enzymes. PLoS ONE. 2012;7. doi:10.1371/journal.pone.0038979

51. Nagao C, Nagano N, Mizuguchi K. Prediction of Detailed Enzyme Functions and Identification of Specificity Determining Residues by Random Forests. PLOS ONE. 2014;9: e84623. doi:10.1371/journal.pone.0084623

52. Noda-García L, Juárez-Vázquez AL, Ávila-Arcos MC, Verduzco-Castro EA, Montero-Morán G, Gaytán P, et al. Insights into the evolution of enzyme substrate promiscuity after the discovery of $\beta\alpha_8$ isomerase evolutionary intermediates from a diverse metagenome. BMC Evolutionary Biology. 2015;15. doi:10.1186/s12862-015-0378-1

53. Garcia-Seisdedos H, Ibarra-Molero B, Sanchez-Ruiz JM. Probing the Mutational Interplay between Primary and Promiscuous Protein Functions: A Computational-Experimental Approach. PLOS Computational Biology. 2012;8: e1002558. doi:10.1371/journal.pcbi.1002558

54. Nesvizhskii AI, Vitek O, Aebersold R. Analysis and validation of proteomic data generated by tandem mass spectrometry. Nature Methods. 2007;4: 787–797. doi:10.1038/nmeth1088

55. Campbell I. Biophysical Techniques - Paperback - Iain D. Campbell - Oxford University Press [Internet]. 2012. Available: `https://global.oup.com/ushe/product/biophysical-techniques-9780199642144?cc=mx&lang=en&`

56. Yang JY, Sanchez LM, Rath CM, Liu X, Boudreau PD, Bruns N, et al. Molecular Networking as a Dereplication Strategy. Journal of Natural Products. 2013;76: 1686–1699. doi:10.1021/np400413s

57. Köcher T, Superti-Furga G. Mass spectrometry–based functional proteomics: From molecular machines to protein networks. Nature Methods. 2007;4: 807–815. doi:10.1038/nmeth1093

58. James LC, Tawfik DS. Conformational diversity and protein evolution – a 60-year-old hypothesis revisited. Trends in Biochemical Sciences. 2003;28: 361–368.

doi:10.1016/S0968-0004(03)00135-X

59. Parisi G, Zea DJ, Monzon AM, Marino-Buslje C. Conformational diversity and the emergence of sequence signatures during evolution. Current Opinion in Structural Biology. 2015;32: 58–65. doi:10.1016/j.sbi.2015.02.005

60. Javier Zea D, Miguel Monzon A, Fornasari MS, Marino-Buslje C, Parisi G. Protein Conformational Diversity Correlates with Evolutionary Rate. Molecular Biology and Evolution. 2013;30: 1500–1503. doi:10.1093/molbev/mst065

61. Gatti-Lafranconi P, Hollfelder F. Flexibility and Reactivity in Promiscuous Enzymes. ChemBioChem. 2013;14: 285–292. doi:10.1002/cbic.201200628

62. 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. 2016;8: 1906–1916. doi:10.1093/gbe/evw125

63. Li L, Stoeckert CJ, Roos DS. OrthoMCL: Identification of Ortholog Groups for Eukaryotic Genomes. Genome Research. 2003;13: 2178–2189. doi:10.1101/gr.1224503

64. Waterhouse RM, Tegenfeldt F, Li J, Zdobnov EM, Kriventseva EV. OrthoDB: A hierarchical catalog of animal, fungal and bacterial orthologs. Nucleic Acids Research. 2013;41: D358–D365. doi:10.1093/nar/gks1116

65. Gao B, Gupta RS. Phylogenetic Framework and Molecular Signatures for the Main Clades of the Phylum Actinobacteria. Microbiology and Molecular Biology Reviews : MMBR. 2012;76: 66–112. doi:10.1128/MMBR.05011-11

66. Sen A, Daubin V, Abrouk D, Gifford I, Berry AM, Normand P. Phylogeny of the class Actinobacteria revisited in the light of complete genomes. The orders "Frankiales" and Micrococcales should be split into coherent entities: Proposal of Frankiales ord. nov., Geodermatophilales ord. nov., Acidothermales ord. nov. and Nakamurellales ord. nov. International Journal of Systematic and Evolutionary Microbiology. 2014;64: 3821–3832. doi:10.1099/ijs.0.063966-0

67. Zhou Z, Gu J, Li Y-Q, Wang Y. Genome plasticity and systems evolution in Streptomyces. BMC Bioinformatics. 2012;13: S8. doi:10.1186/1471-2105-13-S10-S8

68. Kim J-N, Kim Y, Jeong Y, Roe J-H, Kim B-G, Cho B-K. Comparative Genomics Reveals the Core and Accessory Genomes of Streptomyces Species. Journal of Microbiology and Biotechnology. 2015;25: 1599–1605. doi:10.4014/jmb.1504.04008

69. Nam H, Lewis NE, Lerman JA, Lee D-H, Chang RL, Kim D, et al. Network Context and Selection in the Evolution to Enzyme Specificity. Science. 2012;337:

1101–1104. doi:10.1126/science.1216861

70. Copley SD. An Evolutionary Biochemist's Perspective on Promiscuity. Trends in biochemical sciences. 2015;40: 72–78. doi:10.1016/j.tibs.2014.12.004

71. Divergent Evolution of Enzymatic Function: Mechanistically Diverse Superfamilies and Functionally Distinct Suprafamilies. Annual Review of Biochemistry. 2001;70: 209–246. doi:10.1146/annurev.biochem.70.1.209

72. Huang R, Hippauf F, Rohrbeck D, Haustein M, Wenke K, Feike J, et al. Enzyme functional evolution through improved catalysis of ancestrally nonpreferred substrates. Proceedings of the National Academy of Sciences. 2012;109: 2966–2971. doi:10.1073/pnas.1019605109

73. Fondi M, Emiliani G, Liò P, Gribaldo S, Fani R. The evolution of histidine biosynthesis in archaea: Insights into the his genes structure and organization in LUCA. Journal of Molecular Evolution. 2009;69: 512–526. doi:10.1007/s00239-009-9286-6

74. Merino E, Jensen RA, Yanofsky C. Evolution of bacterial trp operons and their regulation. Current opinion in microbiology. 2008;11: 78–86. doi:10.1016/j.mib.2008.02.005

75. Verduzco-Castro EA, Michalska K, Endres M, Juárez-Vazquez AL, Noda-García L, Chang C, et al. Co-occurrence of analogous enzymes determines evolution of a novel $\beta\alpha_8$-isomerase sub-family after non-conserved mutations in flexible loop. Biochemical Journal. 2016;473: 1141–1152. doi:10.1042/BJ20151271

76. Noda-Garcia L. Estudio de la evolución molecular de la función enzimática susando como modelo una enzima con características ancestrales. PhD thesis, Langebio, CINVESTAV. 2012.

77. Petrenko R, Meller J. Molecular Dynamics. eLS. John Wiley & Sons, Ltd; 2001. Available: http://onlinelibrary.wiley.com/doi/10.1002/9780470015902.a0003048.pub2/abstract

78. Molecular Modeling of Proteins Andreas Kukol Springer [Internet]. Available: http://www.springer.com/us/book/9781588298645

79. Sikosek T, Chan HS. Biophysics of protein evolution and evolutionary protein biophysics. Journal of The Royal Society Interface. 2014;11: 20140419. doi:10.1098/rsif.2014.0419

80. Zhou R. Replica Exchange Molecular Dynamics Method for Protein Folding Simulation. In: Bai Y, Nussinov R, editors. Protein Folding Protocols. Humana Press; 2006. pp. 205–223. Available: http://dx.doi.org/10.1385/1-59745-189-4%3A205

81. Bisswanger H. General Aspects of Enzyme Analysis. Practical Enzymology. Wiley-VCH Verlag GmbH & Co. KGaA; 2011. pp. 5–91. Available: http:

//onlinelibrary.wiley.com/doi/10.1002/9783527659227.ch2/summary

82. Hommel U, Eberhard M, Kirschner K. Phosphoribosyl Anthranilate Isomerase Catalyzes a Reversible Amadori Reaction. Biochemistry. 1995;34: 5429–5439. doi:10.1021/bi00016a014

83. Scheer M, Grote A, Chang A, Schomburg I, Munaretto C, Rother M, et al. BRENDA, the enzyme information system in 2011. Nucleic Acids Research. 2011;39: D670–D676. doi:10.1093/nar/gkq1089

84. Van Der Spoel D, Lindahl E, Hess B, Groenhof G, Mark AE, Berendsen HJC. GROMACS: Fast, flexible, and free. Journal of Computational Chemistry. 2005;26: 1701–1718. doi:10.1002/jcc.20291

85. Odokonyero D, Sakai A, Patskovsky Y, Malashkevich VN, Fedorov AA, Bonanno JB, et al. Loss of quaternary structure is associated with rapid sequence divergence in the OSBS family. Proceedings of the National Academy of Sciences of the United States of America. 2014;111: 8535–8540. doi:10.1073/pnas.1318703111

86. Osbourn A. Gene Clusters for Secondary Metabolic Pathways: An Emerging Theme in Plant Biology. Plant Physiology. 2010;154: 531–535. doi:10.1104/pp.110.161315

87. Makarova KS, Aravind L, Galperin MY, Grishin NV, Tatusov RL, Wolf YI, et al. Comparative Genomics of the Archaea (Euryarchaeota): Evolution of Conserved Protein Families, the Stable Core, and the Variable Shell. Genome Research. 1999;9: 608–628. doi:10.1101/gr.9.7.608

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. 2015;14: 111–129. doi:10.1038/nrd4510

89. 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. 2012;194: 855–865. doi:10.1128/JB.06040-11

90. Seitz KW, Lazar CS, Hinrichs K-U, Teske AP, Baker BJ. Genomic reconstruction of a novel, deeply branched sediment archaeal phylum with pathways for acetogenesis and sulfur reduction. The ISME Journal. 2016;10: 1696–1705. doi:10.1038/ismej.2015.233

91. Moustafa A, Loram JE, Hackett JD, Anderson DM, Plumley FG, Bhattacharya D. Origin of Saxitoxin Biosynthetic Genes in Cyanobacteria. PLOS ONE. 2009;4: e5758. doi:10.1371/journal.pone.0005758

92. Medema MH, Osbourn A. Computational genomic identification and functional reconstitution of plant natural product biosynthetic pathways. Natural Product

Reports. 2016;33: 951–962. doi:10.1039/c6np00035e

93. Medema MH, Kottmann R, Yilmaz P, Cummings M, Biggins JB, Blin K, et al. Minimum Information about a Biosynthetic Gene cluster. Nature Chemical Biology. 2015;11: 625–631. doi:10.1038/nchembio.1890

94. Iqbal HA, Low-Beinart L, Obiajulu JU, Brady SF. Natural Product Discovery through Improved Functional Metagenomics in Streptomyces. Journal of the American Chemical Society. 2016;138: 9341–9344. doi:10.1021/jacs.6b02921

95. Ulas T, Riemer SA, Zaparty M, Siebers B, Schomburg D. Genome-Scale Reconstruction and Analysis of the Metabolic Network in the Hyperthermophilic Archaeon Sulfolobus Solfataricus. PLoS ONE. 2012;7. doi:10.1371/journal.pone.0043401

96. Charlesworth JC, Burns BP. Untapped Resources: Biotechnological Potential of Peptides and Secondary Metabolites in Archaea. Archaea. 2015;2015: e282035. doi:10.1155/2015/282035

97. Computational Pan-Genomics Consortium. Computational pan-genomics: Status, promises and challenges. Briefings in Bioinformatics. 2016; doi:10.1093/bib/bbw089

98. Chan C, Jayasekera S, Kao B, Páramo M, Grotthuss M von, Ranz JM. Remodelling of a homeobox gene cluster by multiple independent gene reunions in Drosophila. Nature Communications. 2015;6: 6509. doi:10.1038/ncomms7509

99. Rinke C, Schwientek P, Sczyrba A, Ivanova NN, Anderson IJ, Cheng J-F, et al. Insights into the phylogeny and coding potential of microbial dark matter. Nature. 2013;499: 431–437. doi:10.1038/nature12352

100. Castelle CJ, Wrighton KC, Thomas BC, Hug LA, Brown CT, Wilkins MJ, et al. Genomic Expansion of Domain Archaea Highlights Roles for Organisms from New Phyla in Anaerobic Carbon Cycling. Current Biology. 2015;25: 690–701. doi:10.1016/j.cub.2015.01.014

101. Spang A, Saw JH, Jørgensen SL, Zaremba-Niedzwiedzka K, Martijn J, Lind AE, et al. Complex archaea that bridge the gap between prokaryotes and eukaryotes. Nature. 2015;521: 173–179. doi:10.1038/nature14447

102. Koonin EV. Archaeal ancestors of eukaryotes: Not so elusive any more. BMC Biology. 2015;13. doi:10.1186/s12915-015-0194-5

103. Chaudhari NM, Gupta VK, Dutta C. BPGA- an ultra-fast pan-genome analysis pipeline. Scientific Reports. 2016;6: 24373. doi:10.1038/srep24373

104. Glass JI, Assad-Garcia N, Alperovich N, Yooseph S, Lewis MR, Maruf M, et al. Essential genes of a minimal bacterium. Proceedings of the National Academy of Sciences of the United States of America. 2006;103: 425–430.

doi:10.1073/pnas.0510013103

105. Narechania A, Baker RH, Sit R, Kolokotronis S-O, DeSalle R, Planet PJ. Random Addition Concatenation Analysis: A Novel Approach to the Exploration of Phylogenomic Signal Reveals Strong Agreement between Core and Shell Genomic Partitions in the Cyanobacteria. Genome Biology and Evolution. 2012;4: 30–43. doi:10.1093/gbe/evr121

106. Stamatakis A. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 2014;30: 1312–1313. doi:10.1093/bioinformatics/btu033

107. Powerful tree graphics with ggplot2 [Internet]. Available: `http://joey711.github.io/phyloseq/plot_tree-examples.html`

108. Zacharia VM, Traxler MF. Exploring new horizons. eLife. 2017;6: e23624. doi:10.7554/eLife.23624

109. Woese C. The universal ancestor. Proceedings of the National Academy of Sciences. 1998;95: 6854–6859. Available: `http://www.pnas.org/content/95/12/6854`

110. Woese CR, Gupta R. Are archaebacteria merely derived "prokaryotes"? Nature. 1981;289: 95–96. doi:10.1038/289095a0

111. Woese CR, Kandler O, Wheelis ML. Towards a natural system of organisms: Proposal for the domains Archaea, Bacteria, and Eucarya. Proceedings of the National Academy of Sciences of the United States of America. 1990;87: 4576–4579.

112. Woese CR, Fox GE. Phylogenetic structure of the prokaryotic domain: The primary kingdoms. Proceedings of the National Academy of Sciences of the United States of America. 1977;74: 5088–5090. Available: `http://www.ncbi.nlm.nih.gov/pmc/articles/PMC432104/`

113. Woese CR. There must be a prokaryote somewhere: Microbiology's search for itself. Microbiological Reviews. 1994;58: 1–9. Available: `http://mmbr.asm.org/content/58/1/1`

114. Graham DE, Overbeek R, Olsen GJ, Woese CR. An archaeal genomic signature. Proceedings of the National Academy of Sciences. 2000;97: 3304–3308. doi:10.1073/pnas.97.7.3304

115. Howland JL. The surprising archaea: Discovering another domain of life. New York: Oxford University; 2000.

116. Xu Y, Gogarten JP. Computational Methods for Understanding Bacterial and Archaeal Genomes. World Scientific; 2008.

117. Garrett RA, Klenk H-P. Archaea: Evolution, Physiology, and Molecular Biology.

John Wiley & Sons; 2008.

118. Koonin EV, Mushegian AR, Galperin MY, Walker DR. Comparison of archaeal and bacterial genomes: Computer analysis of protein sequences predicts novel functions and suggests a chimeric origin for the archaea. Molecular Microbiology. 1997;25: 619–637. doi:10.1046/j.1365-2958.1997.4821861.x

119. Koonin EV, Wolf YI. Genomics of bacteria and archaea: The emerging dynamic view of the prokaryotic world. Nucleic Acids Research. 2008;36: 6688–6719. doi:10.1093/nar/gkn668

120. Koonin EV. The Turbulent Network Dynamics of Microbial Evolution and the Statistical Tree of Life. Journal of Molecular Evolution. 2015;80: 244–250. doi:10.1007/s00239-015-9679-7

121. Land M, Hauser L, Jun S-R, Nookaew I, Leuze MR, Ahn T-H, et al. Insights from 20 years of bacterial genome sequencing. Functional & Integrative Genomics. 2015;15: 141–161. doi:10.1007/s10142-015-0433-4

122. Nishida H. Evolution of genome base composition and genome size in bacteria. Frontiers in Microbiology. 2012;3. doi:10.3389/fmicb.2012.00420

123. Coyle M, Hu J, Gartner Z. Mysteries in a Minimal Genome. ACS Central Science. 2016;2: 274–277. doi:10.1021/acscentsci.6b00110

124. O'Meara B. CRAN Task View: Phylogenetics, Especially Comparative Methods. 2016; Available: `https://CRAN.R-project.org/view=Phylogenetics`

125. Larsson J, Nylander JA, Bergman B. Genome fluctuations in cyanobacteria reflect evolutionary, developmental and adaptive traits. BMC Evolutionary Biology. 2011;11: 187. doi:10.1186/1471-2148-11-187

126. Whitton BA. Ecology of Cyanobacteria II: Their Diversity in Space and Time. Springer Science & Business Media; 2012.

127. Cohen GN. The biosynthesis of histidine and its regulation. Microbial Biochemistry. Springer Netherlands; 2004. pp. 225–230. Available: `http://link.springer.com/chapter/10.1007/978-1-4020-2237-1_29`

128. Plach MG, Reisinger B, Sterner R, Merkl R. Long-Term Persistence of Bifunctionality Contributes to the Robustness of Microbial Life through Exaptation. PLOS Genetics. 2016;12: e1005836. doi:10.1371/journal.pgen.1005836

129. Battistuzzi FU, Feijao A, Hedges SB. A genomic timescale of prokaryote evolution: Insights into the origin of methanogenesis, phototrophy, and the colonization of land. BMC Evolutionary Biology. 2004;4: 44. doi:10.1186/1471-2148-4-44

130. Lapierre P, Gogarten JP. Estimating the size of the bacterial pan-genome. Trends in Genetics. 2009;25: 107–110. doi:10.1016/j.tig.2008.12.004

131. Větrovský T, Baldrian P. The Variability of the 16S rRNA Gene in Bacterial

Genomes and Its Consequences for Bacterial Community Analyses. PLoS ONE. 2013;8. doi:10.1371/journal.pone.0057923

132. Brettin T, Davis JJ, Disz T, Edwards RA, Gerdes S, Olsen GJ, et al. RASTtk: A modular and extensible implementation of the RAST algorithm for building custom annotation pipelines and annotating batches of genomes. Scientific Reports. 2015;5: 8365. doi:10.1038/srep08365

133. chesterismay. Updated R Markdown thesis template [Internet]. Chester's R blog. 2016. Available: `https://chesterismay.wordpress.com/2016/09/01/updated-r-markdown-thesis-template/`

134. Barona-Gómez F, Cruz-Morales P, Noda-García L. What can genome-scale metabolic network reconstructions do for prokaryotic systematics? Antonie van Leeuwenhoek. 2012;101: 35–43. doi:10.1007/s10482-011-9655-1

135. Weber T, Blin K, Duddela S, Krug D, Kim HU, Bruccoleri R, et al. antiSMASH 3.0—a comprehensive resource for the genome mining of biosynthetic gene clusters. Nucleic Acids Research. 2015;43: W237–W243. doi:10.1093/nar/gkv437

136. Medema MH, Blin K, Cimermancic P, Jager V de, Zakrzewski P, Fischbach MA, et al. antiSMASH: Rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences. Nucleic Acids Research. 2011;39: W339–W346. doi:10.1093/nar/gkr466

137. Molina ST, Borkovec TD. The Penn State worry questionnaire: Psychometric properties and associated characteristics. In: Davey GCL, Tallis F, editors. Worrying: Perspectives on theory, assessment and treatment. New York: Wiley; 1994. pp. 265–283.

138. Reed College. LaTeX your document [Internet]. 2007. Available: `http://web.reed.edu/cis/help/LaTeX/index.html`

139. Noble SG. Turning images into simple line-art. Undergraduate thesis, Reed College. 2002.

140. Angel E. Interactive computer graphics : A top-down approach with opengl. Boston, MA: Addison Wesley Longman; 2000.

141. Angel E. Batch-file computer graphics : A bottom-up approach with quicktime. Boston, MA: Wesley Addison Longman; 2001.

142. Angel E. Test second book by angel. Boston, MA: Wesley Addison Longman; 2001.