

CORASON

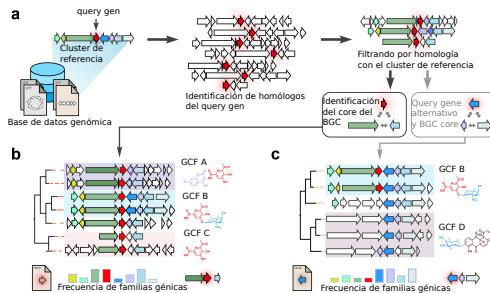


Figure 1: La herramienta corason localiza familias de clusters biosintéticos en un linaje genómico partiendo de un cluster y un gen de referencia. Todos los contextos genómicos que contengan ese gen y algún otro gen del cluster de referencia serán encontrados en el linaje seleccionado por el usuario. CORASON identifica el core génico de la familia. La información del core se utiliza para organizar filogenéticamente todos los miembros de la familia del BGC, es decir todas las variantes del BGC serán organizadas. El core génico está relacionado con el core de la molécula, la parte variable del BGC codifica enzimas accesoriales que producen ornamentos i.e. variantes moleculares. Al cambiar de gen de referencia CORASON permite explorar otras familias de BGC que contengan las mismas modificaciones. Los resultados se presentan en una visualización que permite al mismo tiempo apreciar variación a nivel de presencia-ausencia de genes entre miembros de una familia de BGCs, como también apreciar variación a nivel de secuencia a través de un gradiente de color entre genes conservados entre una variante y el BGC de referencia.

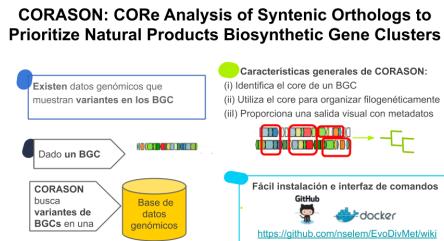
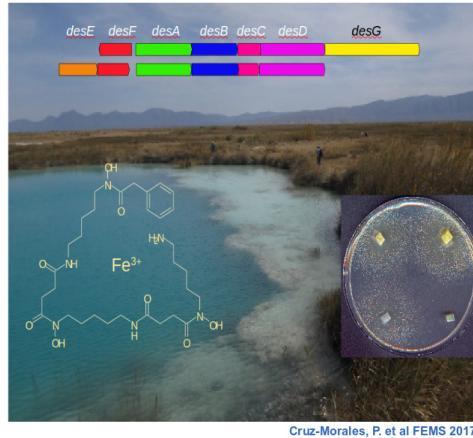


Figure 2: CORASON es una herramienta en línea de comandos para organizar filogenéticamente la variación de una familia de clusters. Dada la existencia de variantes de BGCs en bases de datos de linajes genómicos se diseñó CORASON. Corason se ejecuta en línea de comandos, su función es identificar el core génico de un BGC y utilizarlo para presentar una visualización de las variantes de la familia organizadas filogenéticamente

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

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

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



Cruz-Morales, P. et al FEMS 2017

Figure 3: Variantes del cluster biosintético de desferroxiamina fueron identificadas por CORASON en Actinobacterias de cuatro ciéngas. Una variante del BGC de desferroxiamina fue identificada. Esta variante se diferencia del cluster reportado por la ganancia de una penicilin amidasa. A este gen extra se le llamó *desG*. Fue verificado que la variación génica está ligada con la variación molecular

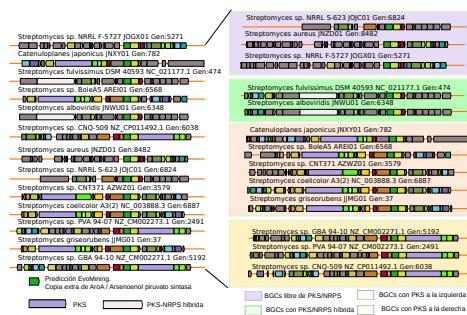


Figure 4: Contexto genómico conservado en las expansiones de AroA. La arsenoenol piruvato sintasa fue descubierta en un árbol de EvoMining como parte de una rama de expansiones de AroA en Actinobacteria. El contexto genómico de la arsenoenol piruvato sintasa de *S. coelicolor* tiene un core conservado en Actinobacteria. Este BGC está dedicado a la síntesis de metabolitos secundarios de tipo arsenolípidos. Primero se identificaron en otras secuencias genómicas contextos que contengan el homólogo de AroA y algún otro gen de su vecindad en *S. coelicolor*. A continuación, se ordenaron manualmente los contextos obtenidos y se pudo identificar al menos cuatro diferentes clases de BGCs. La primera clase mostrada en un rectángulo morado no contiene PKS o NRPS, la segunda enmarcada en verde contiene una PKS-NRPS híbrida. La tercera clase incluye una PKS a la izquierda y a más de cinco genes de distancia de la Arsenoenol piruvato sintasa y finalmente la última clase contiene una PKS a la derecha y a sólo un gen de distancia de esta enzima. Esta figura muestra que existen variantes de un BGC



Figure 5: Azoxy BGC

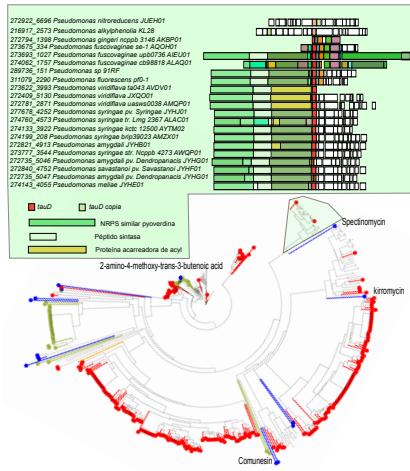


Figure 6: En *Pseudomonas* el gen *tauD* tiene un contexto que incluye genes de metabolismo especializado. Variantes de este contexto están distribuidas en varios organismos. En particular

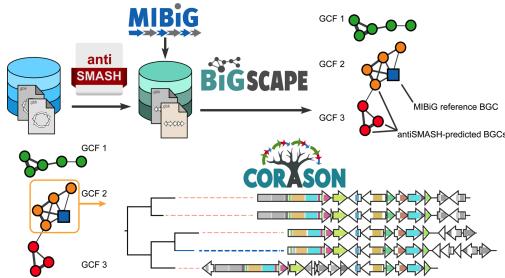


Figure 7:

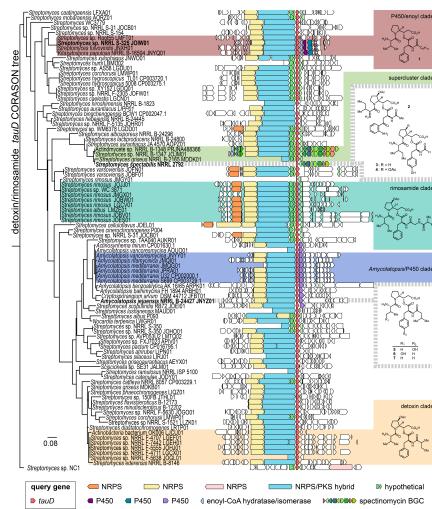


Figure 8: n

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

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

Algoritmo y características de CORASON

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

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

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

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

Así pues, cuando se habla de un BGC no se debe pensar que este es igual en todos los organismos del linaje, es decir que tiene el mismo contenido génico. Hay variación tanto a nivel de contenido génico como a nivel de secuencia entre los genes en común. Esta variación produce la promiscuidad de producto, una familia de BGCs produce distintos productos a partir de los mismos precursores. En este trabajo vamos a tomar como BGC de referencia a los anotados en MIBiG, que son de los que se tienen datos experimentales respecto al producto reportado. Todas sus variantes que contengan al menos dos genes en común, uno de referencia seleccionado por el usuario y otro cualquiera pero común con el BGC de referencia son considerados parte de la familia del BGC.

Como ejemplo de familias de BGCs podemos pensar a los operones. Un operón es el conjunto de genes dedicados a la síntesis de algún metabolito del metabolismo central, estos genes suelen encontrarse juntos en el cromosoma y transcribirse simultáneamente. Ejemplos de estos operones son los que producen los aminoácidos histidina y triptófano. Estos “BGCs” hace millones de años fueron posiblemente parte del metabolismo especializado y debido a su éxito se fijaron en lo que ahora vemos como BGCs conservados en ciertos linajes genómicos. En estos casos las fronteras son claras y la variación génica es poca. Aún así existen otros ejemplos donde puede constatarse variación en las rutas de síntesis de mecanismos centrales de

metabolismo procariota. En el caso de histidina y triptófano, como hablaremos en el siguiente capítulo, es la variación a nivel de secuencia y no a nivel de composición génica la que produce diversidad de producto.

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

Aplicaciones de CORASON en Actinobacteria y Pseudomonas

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

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

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

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

Estas familias pueden clasificarse

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

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

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

En el linaje de *Streptomyces* EvoMining obtuvo como predicción una arsenoenol piruvato sintasa en una rama del árbol de expansiones de la familia 3-fosfoshikimato-1-carboxivinyl transferasa (AroA). Estudios de mutagénesis y de expresión diferencial génica en presencia de arsénico confirmaron que en *Streptomyces coelicolor* y en *Streptomyces lividans* esta enzima pertenece a un BGC que sintetiza arsenolípidos [Pablo tesis]. El contexto genómico de AroA en estos organismos incluye una enzima PKS localizada a seis genes de distancia. AntiSMASH predice los BGCs tipo PKS de estas enzimas pero no incluye en ellos a la arsenoenol piruvato sintasa. El valor de EvoMining fue descubrir que esta copia extra de AroA estaba dedicada al metabolismo especializado y por tanto bien podía tener su propio BGC o dada la cercanía con las PKS podía ser parte de estos PKS-BGCs. Esta ruta de síntesis de arseno compuestos, fue descubierta a partir de una rama de copias extra de AroA donde se apreciaba diversidad en las secuencias de aminoácidos. Una pregunta posible es si esta diversidad de secuencia a nivel de enzima ascendía a un siguiente nivel. Es decir si también existe diversidad a nivel del BGC, si se encuentran variantes con distintos patrones de presencia/ausencia en los genes que coponen al BGC de *S coelicolor*. O más aun, quedaba por investigar si el BGC tenía cierto grado de conservación o era exclusivo de estos dos organismos *S. coelicolor* y *S. lividans*.

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

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

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

CORASON analiza la conservación del contexto genómico de los EvoMining hits

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

Figura [CorasonAzoxy] Azoxy BGC

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

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

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

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

[CorasonBiGSCAPE] texto texto

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

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

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

Figura [TauDActinobacteria]texto

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

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

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

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

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

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

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

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

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

Los tiempos de retención de ionies y los patrones de fragmentación de los últimos dos compuestos también fueron observados en extractos de *Streptomyces* sp. NRRL B-1347 parte del clado del supercluster, confirmando la habilidad de CORASON para guiar un descubrimiento, mediante la utilización de la filogenia a pesar de lo

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

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

El tercer clado que se estudió del super clan al que pertenecen las familias rimosamide-detoxin, contiene BGCs casi enteramente provenientes del género Amycolatopsis. Este clado está marcado en morado en la [Fig TauDActinobacteria]. Este clado de BGCs también contiene un gen P450 único entre los BGCs del árbol, así que fue llamado el clado ‘Amycolatopsis/P450 clade’. Aunque no se contaba con datos metabolómicos de las cepas del clado de BGCs definidos por BiG-SCAPE como una gen cluster family (GCF), la visualización filogenética de CORASON permitió la selección de una cepa de Amycolatopsis de la que se tenían datos metabolómicos con un BGC muy similar, y que también contiene el gen P450 ([Fig TauDActinobacteria], línea gris cerca del clado *Amycolatopsis/P450*). Análisis de datos de tandem MS de extracto fermentado de *Amycolatopsis jejuensis* NRRL B-24427 reveló isómeros de detoxins P1 (5; [Fig TauDActinobacteria], S15) que contienen tirosina, P2 (6; [Fig TauDActinobacteria], S15) mostrando fenilalanina y una valina hidroxilada, así como la detoxin P3, un análogo cercano libre de hidroxilación (7; [Fig TauDActinobacteria], S15). Una validación de la asignación de aminoácidos observados en los patrones de fragmentación de MS/MS se consiguió mediante el uso de experimentos de incorporación de isotopos estables etiquetados de aminoácidos (Figs. S33–S36, S38–S41, and S43–S44). ## CORASON permitió expandir las capacidades de EvoMining y explorar la promiscuidad de familias de BGCs de enzimas divergentes de metabolismo central. Nuestros resultados ilustran como BiG-SCAPE puede identificar conjuntos de BGCs relacionados, en un gran número de secuencias de genomas. Además al usar CORASON para reconstruir las filogenias de BGC para ordenar visualmente la evolución de un cluster biosintético y su diversidad proveen herramientas poderosas para el descubrimiento de nuevos clados de BGC que codifican en consecuencia para nueva química. Respecto a los BGCs detoxin/rimosamide, CORASON mostró habilidad para ayudar a minar bases de datos genómicas y descubrir siete nuevas detoxins. Específicamente, la organización de las variantes de los BGC facilitó la identificación de los correspondientes variaciones en la estructura química -la presencia de un enoyl-CoA hidratasa/isomerasa corresponde a la familia de amida ácido graso detoxin S1 y la presencia de un gen P450 corresponde a la presencia de hidroxilaciones en detoxins P1–P3.

Discusión

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