UNIVERSIDAD DE ANTIOQUIA

Análisis bioinformático del transcriptoma de *Streptomyces clavuligerus* y establecimiento de una red de interacción asociado al metabolismo de ácido clavulánico.

Juan Pablo Martínez Aldana

Universidad de Antioquia

Facultad de Ingeniería

Ingeniería Bioquímica

El Carmen de Viboral, Colombia

2020

Análisis bioinformático del transcriptoma de *Streptomyces clavuligerus* y establecimiento de una red de interacción asociado al metabolismo de ácido clavulánico.

Juan Pablo Martínez Aldana

Informe de investigación
como requisito para optar al título de:
Ingeniero Bioquímico.

Asesores (a): Laura Inés Pinilla Mendoza, PhD.
Co- Director: León Felipe Toro Navarro, PhD.

Universidad de Antioquia
Facultad de Ingeniería
Ingeniería Bioquímica
El Carmen de Viboral, Colombia
2020

Resumen.

El ácido clavulánico es un inhibidor de β-lactamasas, y es producido a nivel industrial por procesos fermentativos, usando la bacteria *Streptomyces clavuligerus***.** La eficiencia de la producción de AC está asociada con la composición de los medios, las condiciones de cultivo y las características fisiológicas y genéticas de la cepa. Sin embargo, gran parte de las vías moleculares que rigen la regulación de AC en *S. clavuligerus* siguen siendo aún desconocidas, y a pesar de numerosos estudios, los mecanismos reguladores que se encuentran presentes en la ruta de biosíntesis de AC aún no se han identificado completamente**.** En este trabajo, se realizó un análisis transcriptomico, usando datos provenientes de RNA-seq de la cepa de *S. clavuligerus* ATCC 27064 la cual fue cultivada en dos medios de cultivo. El primero, era un medio complejo, a base de soja el cual era rico en nutrientes. El segundo medio era químicamente definido, en el cual tenía escasez de nutrientes . En total se encontraron 1505 genes que fueron expresados diferencialmente, 587 genes fueron expresados bajo las condiciones favorables del cultivo de soja y 918 en condiciones de cultivo desfavorables. Además, se usó la herramienta string, disponible en cytoscape, para construir una red de interacción con los datos obtenidos en el análisis de expresión diferencial, usando un puntaje de confidencialidad de 0.7. Se encontraron 524 nodos 243 bordes.Se realizó un análisis de enriquecimiento usando string para identificar cuáles de esos genes, formaban parte en la producción de antibióticos, y otros procesos de síntesis de interés, como lo es la producción de arginina. Se usó la herramienta cluster Maker 2, y se encontró un grupo de genes diferencialmente expresados, los cuales contribuyen a la producción de alanina el cual es un precursor importante para la producción de ácido clavulánico. El análisis de transcriptoma en *S. clavuligerus* contribuye a la identificación de la abundancia de transcripción durante el desarrollo celular y las perturbaciones ambientales, esto nos ayuda entender cómo se comporta el microorganismo bajo diversas condiciones y así poder aportar y entender cómo funciona la producción de ácido clavulánico.

Palabras clave: *Análisis transcriptómico, red de interacción, Ácido clavulánico, Expresión diferencial.*

1. Introducción.

El ácido clavulánico (AC), es un inhibidor de la β-lactamasa, que es producido industrialmente por la fermentación de *Streptomyces clavuligerus* [1]. La eficiencia de la producción de AC está asociada con la composición de los medios, las condiciones de cultivo y las características fisiológicas y genéticas de la cepa[3]. Sin embargo, gran parte de las vías moleculares que rigen la regulación de AC en *S. clavuligerus* siguen siendo aún desconocidas, y a pesar de numerosos estudios, los mecanismos reguladores que se encuentran presentes en la ruta de biosíntesis de AC aún no se han identificado completamente [2] . La producción de AC está controlada por un grupo de genes reguladores, entre los que se encuentran reguladores pleiotrópicos como BldG y BldA [22] y AdpA [23] , y reguladores específicos de vía como lo son CcaR, un activador transcripcional codificado por *ccaR* ubicado en el grupo de genes de cefamicina C y el activador ClaR específico de la ruta AC

codificado por claR [24]. Muchos de los estudios se han centrado en genes individuales [3], donde se encontró que al inactivar el gen *ccaR*, tanto los genes "tempranos" como "tardíos" de la ruta de biosíntesis de AC fueron regulados negativamente, mientras que cuando se hace la inactivación de *claR,* se produjo un bajo nivel de expresión en cluster de genes de tardíos de AC (grupo AC), esto al ser comparadas las cepas modificadas con la cepa silvestre [22,23,24]. Otro estudio encontró que la supresión de *bldA* condujo a la sobreexpresión de las proteínas Cas2, OppA1 y GcaS, mientras que la eliminación de *bldG* resultó en la baja formación de Bls2 [25].

Para poder desarrollar un enfoque global y mejorar el entendimiento de los sistemas regulatorios implicados en la producción de metabolitos secundarios, a través de la ingeniería genética es necesario utilizar métodos modernos con el fin de obtener información sobre la expresión temporal y condicional de factores reguladores globales, reguladores de vías específicas y tasas limitadas de producción de enzimas. Gracias al desarrollo de herramientas como RNA-Seq, la cual ha demostrado ser particularmente útil en el estudio del rendimiento metabólico celular a través del entendimiento de reguladores transcripcionales y análisis de expresión génica diferencial, se ha logrado contribuir al mejoramiento de las cepas [6]. Como resultado, se ha logrado una mejor comprensión de las vías metabólicas y los mecanismos reguladores involucrados en la biosíntesis de metabolitos importantes, como el AC.

Con la información proporcionada por el análisis transcripcional, se puede desarrollar un modelo de análisis de datos, con el cual es posible generar un estudio de red que proporciona un enfoque más productivo para la visualización de los datos, con esto se pueden llevar la información a un marco gráfico (red), lo que presenta ventajas ya que permite la adopción de técnicas desarrolladas en teoría de gráficos, ingeniería y ciencias de la computación como lo es la bioinformática. Estos enfoques pueden relacionar directamente interacciones biológicas específicas; siendo el análisis de red de co-expresión de genes un método utilizado para explorar las complejas relaciones entre genes y fenotipo [11,12]. Con este proyecto, se planea realizar un análisis bioinformático a partir del transcriptoma de *S. clavuligerus* para el desarrollo de una red de interacción génica que permita contribuir a la comprensión de la biosíntesis de algunos metabolitos secundarios y posiblemente algunos mecanismos regulatorios que podrían estar involucrados en la síntesis de A.

2. Objetivos:

2.1 Objetivo General:

- Proponer a partir de datos de expresión génica de *Streptomyces clavuligerus*, cultivada en diferentes condiciones de medio de cultivo, una red de interacción génica que permita entender posibles patrones regulatorios asociado principalmente al metabolismo de ácido clavulánico.

2.2 Objetivos Específicos:

- Realizar un análisis bioinformático de los datos de expresión génica de *Streptomyces clavuligerus,* enfatizando en los genes diferencialmente expresados asociados con la biosíntesis de ácido clavulánico.

- Hacer uso de las redes de interacción disponibles en las bases de datos del género *Streptomyces* sp., y los datos encontrados en el análisis transcriptómico, para proponer una red de coexpresion de genes.

3. Marco Teórico:

Streptomyces clavuligerus (S.clavuligerus) es un actinomiceto productor de una gran variedad de metabolitos secundarios, entre los que se encuentran productos de gran interés a escala industrial como ácido clavulánico (AC), con actividad inhibidora de betalactamasas[3]. Debido a que el AC es un potente inhibidor de β-lactamasas, producidas por microorganismos resistentes de los géneros, *Staphylococcus sp, Shigella sp, Escherichia sp, Klebsiella sp, Salmonella sp y Proteus sp.* [13,14]. La ruta de síntesis de AC inicia con la condensación de los precursores metabólicos D- gliceraldehído-3-fosfato (G3P), precursor C3 y el precursor C5, la L-arginina [3]. El primer paso da lugar a N2- (2-carboxietil) -arginina y es catalizada por la carboxietil agininasintasa (CeaS). La β-lactama sintetasa (Bls) cataliza el segundo paso convirtiendo la carboxietil arginina en ácido desoxi guanidino-proclavamínico, que ahora contiene el anillo de β-lactama. El ácido desoxi guanidino-proclavamínico luego se hidroxila para dar ácido guanidina-proclavamínico por medio de la clavaminate sintetasa (Cas) Posteriormente, el ácido proclavamínico amidino hidrolasa (Pah) cataliza la reacción del ácido guanidina-proclavamínico al ácido proclavamínico, donde se elimina el grupo guanidino del extremo de la molécula derivado de la arginina. La clavaminate sintasa (Cas) cataliza la formación del primer intermedio bicíclico a través del cierre oxidativo del anillo de ácido proclavamínico para dar ácido dihidro claveamínico seguido de desaturación para formar ácido clavamínico. El clavaldehído, el último intermedio conocido de la vía, finalmente se convierte en ácido clavulánico por la deshidrogenasa del ácido clavulánico (Cad) [26].

Por otro lado, en la actualidad, la transcriptómica permite mediciones cuantitativas de la expresión de ARNm y las variaciones entre diferentes estados, lo que refleja los genes que se sobreexpresan o regulan negativamente en momentos y condiciones particulares [3]. Gracias a esto la ruta de biosíntesis de AC es parcialmente conocida [25,26], siendo una de las etapas más estudiadas, la que involucra el gen gcas, también llamado orf 17, el cual convierte el ácido clavamínico en ácido N-glicil clavamínico mediante la expresión de la enzima ácido glicil-clavamínico-sintasa (GCAS), el ácido N-glicil-clavamínico es el primer intermediario biosintético identificado, específico para la producción de AC y no se encuentra presente en la ruta de síntesis de clavamas 5S [11]. En este sentido, el estudio de los perfiles de expresión génica o transcriptoma contenidos en el genoma de *S. clavuligerus*, podría contribuir a la identificación de genes claves expresados durante proceso celular determinado, además de contribuir a los avances en ingeniería metabólica y la mejora de cepas para aumentar la producción de ácido clavulánico, mediante diferentes técnicas de ingeniería metabólica evaluando de manera efectiva, las alteraciones causadas en el proteoma [27]. Es por esto que el análisis transcriptómico, se convierte en una herramienta esencial para interpretar los elementos principales del genoma [15], siendo el fin último de este tipo de análisis estudiar e identificar transcripciones, caracterizar la complejidad estructural de la transcripción además del contenido del codificacion[16]. En cuanto al género de *Streptomyces*, gracias al análisis transcriptómico se ha encontrado múltiples avances, en el estudio de resistencia a antibióticos en *Streptomyces coelicolor* ayudando a determinar los efectos de la ciprofloxacina (CIP), este se investigó realizando una

medición de las proteínas expresadas diferencialmente al someter la bacteria a diferentes concentraciones de (CIP) [28]. En otro estudio diferente a la resistencia microbiana, se encontró el mecanismo a través del cual el género *Streptomyces*, puede ser útil a la hora de la degradación de compuestos altamente contaminantes como lo es el γ-hexaclorociclohexano (lindano) y su posible potencial en biorremediación [29].

Una de las tecnologías más usadas para el análisis del proteoma es la de RNA-Seq, la cual permite estimar la expresión de genes, esta se introdujo por primera vez en 2008 y durante la última década se ha utilizado ampliamente debido a la disminución de los costos y la popularización de los núcleos de secuenciación de recursos compartidos en muchas instituciones de investigación [17], En general, esta tecnología es muy útil para el análisis de expresión diferencial que involucra condiciones específicas [18][3]. Finalmente, la significancia de los datos producidos debe ser evaluada desde un contexto biológico, para darle validez al análisis, y poder mejorar el esclarecimiento de los cambios a nivel de transcriptoma, y representar de una manera más adecuada la forma en que un conjunto de genes interactúan entre sí para formar un módulo funcional y cómo se relacionan los diferentes módulos de genes; podemos hacer uso de un análisis de red de coexpresión ya que este es un método utilizado con frecuencia para explorar las complejas relaciones entre genes y fenotipos, el cual es ampliamente usado para el análisis de datos producidos por secuenciación de alto rendimiento [20]. Los estudios encontrados hasta la fecha se centran en el desarrollo de redes de regulación génica, por ejemplo se encontró un estudio proteómico del retraso diaúxico en la procariota diferenciadora *Streptomyces coelicolor* para encontrar una red reguladora de proteínas inducidas por el estrés y enzimas metabólicas centrales [30]. En otro estudio se usó este tipo de redes, para poder realizar caracterización de Fosfopanteteinil transferasas en *Streptomyces tsukubaensis* L19 [31]. Para el caso de *S. clavurigenus*, no se encontraron muchos estudios en los cuales se hiciera énfasis en este tema, es importante desarrollar mayores investigaciones en cuanto al desarrollo de redes de interacción génica, ya que esto nos permitirá mejorar nuestro entendimiento, en cuanto a los mecanismos reguladores en la producción de antibióticos de una manera más eficiente. Las redes de coexpresión se pueden construir utilizando cytoscape y las herramientas disponibles en este como lo son string y clusterMarker2, ya que este permite cargar los datos de expresión génica y elegir el diversos métodos para inferir y visualizar la red , así como integrar diferentes métodos para obtener un resultado más seguro. Este tipo de análisis representa como un conjunto de genes interactúan entre sí para formar un módulo funcional y cómo se relacionan los diferentes módulos de genes [11].

4. Metodología

1. Revisión bibliográfica.

Se realizó una revisión bibliográfica exhaustiva en las diferentes bases de datos, y demás repositorios bioinformáticos disponibles, además de hacer uso de diversos artículos e investigaciones las cuales se enfocan en este tipo de análisis. Los parámetros estadísticos y los valores máximos permitidos son determinados basados en las condiciones de las muestras según lo recomienda la bibliografía.

2. Análisis de los resultados del RNA-Seq y datos experimentales.

El sistema operativo base del entorno de trabajo bioinformático usado fue linux Ubuntu 20.04 LTS [34]. La bacteria *Streptomyces clavurigenus* y parte de los datos experimentales del transcriptoma que se utilizaron para el presente análisis, fueron extraídos de una tesis doctoral realizada en el grupo de Bioprocesos, de la Facultad de Ingeniería de la Universidad de Antioquia cultivada a diferentes condiciones ambientales, en dos medios; un medio químicamente definido el cual tenía escasez de nutrientes y un medio complejo a base de soja el cual era favorable para la producción de AC.

3. Análisis de calidad y procesamiento de datos:

Se realizó el control de calidad de las lecturas obtenidas mediante la tecnología de Rna-seq, usando el kit de herramienta FastQC (Galaxy Versión 0.72+galaxy1)[27]. Con el fin de mejorar la calidad de las lecturas, la precisión y el rendimiento computacional se usó el paquete de herramientas contenidos en FASTX-Toolkit, de las cuales se hizo uso de Tripomatic, para remover adaptadores y filtrar según la calidad [19].Para la remoción del rRNA ribosomal que las lecturas contenían, se usó la herramienta riboPiker [34].

4. Alineamiento o mapeo.

De los resultados obtenidos después del procesamientos de los datos y el análisis del control de calidad, se procedió a realizar el alineamiento y mapeo de las lecturas cortas. Para este caso se contó con un genoma de referencia disponible en la base de datos del NCBI [3] con el código de acceso ASM169367v1 y el mapeo de las lecturas en el genoma de referencia para . Se va a hacer uso de la herramienta BOWTIE2 [32].

5. Análisis de expresión diferencial (genes diferencialmente expresados, GDEs)

Después de los pasos iniciales de control de calidad, la eliminación de valores atípicos y el filtrado, los datos estarán listos para el análisis de expresión diferencial. Para poder detectar y comparar los cambios en el transcriptoma, a través de las muestras tomadas en diferentes fases del experimento se usó edgeR para evaluar si hay GDEs estadísticamente significativos entre dos grupos usando un modelo binomial negativo [32]. La tabla de datos resultante asignó valores P , ajustando valores (calculados usando el método de tasa de descubrimiento falso [FDR] de Benjamini-Hochberg para ajustar las pruebas de hipótesis múltiples) y registrar 2 veces los cambios para cada gen [19].

6. Creación de la red de interacción génica (RIG).

Finalmente, la significancia de los datos producidos se evaluó desde un contexto biológico. Para darle validez al análisis, y poder mejorar el esclarecimiento de los cambios a nivel de transcriptoma, la creación de una red de interacción génica (RIG) y la visualización de datos de RNA-seq [19]. Durante esta etapa se usarán herramientas disponibles de visualización y analizar múltiples muestras de RNA-seq, como cytoscape la cual cuenta con una herramienta incorporada que da acceso a las bases de datos de string, y otras herramientas como cluster Marker 2 las cuales permiten ampliar el análisis y organizar la red de genes [30]. Con esta actividad se logró encontrar una red de interacción, asociado a la expresión génica de *S. clavuligerus*[35].

5. Materiales y Métodos.

5.1 Recuperación de Datos.

Los datos sin procesar de Rna-seq fueron proporcionados por el grupo de bioprocesos de la universidad de antioquia, obtenidos mediante el desarrollo de una tesis doctoral desarrollada en el grupo[3] .Los datos corresponden al transcriptoma de *streptomyces clavuligerus* ATCC 27064, la

cual fue cultivada bajo condiciones ambientales diferentes. La primera fue un medio de cultivo basado en soja el cual era favorable para la producción de AC. El segundo medio era químicamente definido. con escasez de nutrientes. El conjunto de datos contiene dos lecturas para cada condición de cultivo, Con_Ac1, Con_Ac2 para las condiciones favorables de cultivo , Sin_Ac1 y Sin_Ac2 para el medio químicamente definido.

5.2 Análisis De los Datos

El análisis de calidad de los datos, se realizó usando el kit de herramientas de FasQC (Versión Galaxy 0.72+galaxy1)[27]. Los adaptadores y las lecturas de baja calidad se recortaron con Trimmomatic (Versión de Galaxy 0.38.0)[19]. Para el mapeo de las lecturas, se usó BOWTIE2 (Versión de Galaxy 2.3.4.3+galaxy0) [32]. usando el genoma de referencia disponible en la base de datos del NCBI [3] ATCC27064, disponible en Genbank (número de acceso NZ_CM000913.1 y NZ_CM000914.1 , cromosoma y megaplasmido, respectivamente) [32]. La expresión de cada gen se evaluó usando la herramienta HTseq−count (Versión de Galaxy 0.9.1). Los recuentos entregados, se normalizaron y se utilizaron para las pruebas de expresión diferencial usando EdgeR (Versión de Galaxy 3.24.1+Galaxy1) [32]. Los genes con un valor de log2FC > 1.0 y un valor−p ajustado < 0.05 fueron considerados como sobre regulados, y de modo contrario, aquellos con un log2FC < −1.0 and padj < 0.05 fueron considerados sub expresados.

5.3 Red de Co-expression

La red de Co-expresión se reconstruyó por aparte para cada conjunto de datos, se usaron los valores de logFC obtenidos de EdgeR. La visualización de la red, se hizo mediante la herramienta Cytoscape v3.8.0 [39]. Se usó la base de datos de STRING, la cual se encuentra incluida Cytoscape. La agrupación de la red. se realizó usando el algoritmo de agrupamiento de Markov (MCL). en ClusterMaker2 (Morris et al., 2011).

6. Resultados.

6.1 Procesamiento de datos e identificación de expresión diferencial.

El objetivo de este estudio es encontrar qué genes y las redes provenientes de ellos se encuentran involucrados en la biosíntesis de AC, en consecuencia con esto se compararon los perfiles de transcripción de dos muestras, de *Streptomyces clavuligerus.* El informe FastQC muestra si existe contaminación por parte de los adaptadores, además permite conocer la distribución de GC, k-meros

sobrerrepresentados y lecturas duplicadas para detectar errores de las secuencias. Para este caso se encontró que la calidad de los datos era buena, ya que se encontraban en el percentil superior, como se muestra en la figura 1. De los resultados arrojados por FastQC, se analizó la calidad media de los datos. la cual está representada en la figura 1[a],1[b],1[c] y 1[d]. El eje *y* muestra los puntajes de calidad, cuanto mayor sea el puntaje, mejor será la calidad de la base. Este eje se divide en tres zonas, una de muy buena calidad (verde), de calidad razonable (naranja) y llamadas de baja calidad (rojo)[29]

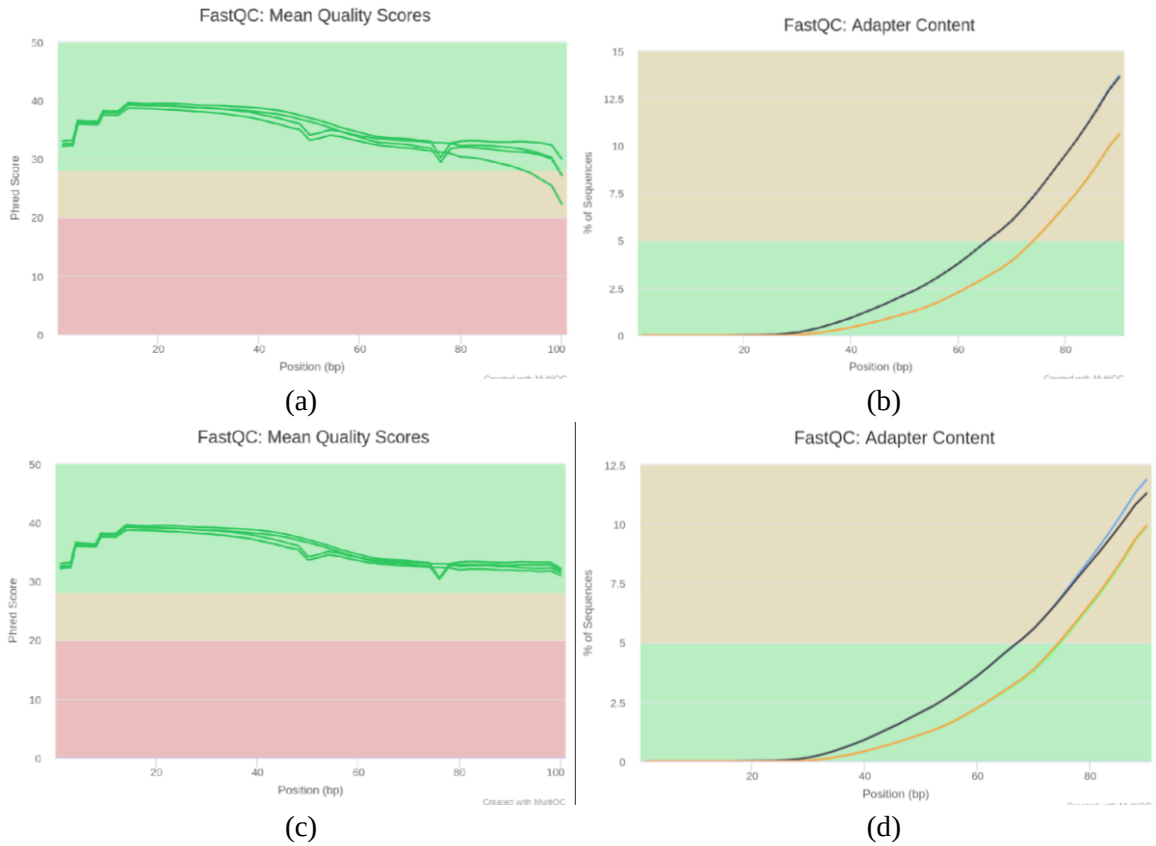

Figura 1: Comparación de la calidad de las lecturas, sin realizar ningún tipo de tratamiento, a y b y haciendo uso de la herramienta timmomatic, para la eliminación de secuencias cortas y de adaptadores c y d. En la parte a y b se muestra la calidad media de las secuencias, haciendo uso del valor de calidad de phred en la parte b y d se hace referencia al contenido de adaptadores en las secuencias.

La puntuación media sobre la secuencia está cayendo al final de las secuencias. Esto se debe a que las tecnologías de secuenciamiento en este caso Rna-seq, no son perfectas, y tienden a incorporar nucleótidos incorrectos hacia el final del proceso.[2] Otro aspecto importante a tener en cuenta en la calidad de las lecturas, es el contenido de adaptadores. que se utilizaron para la secuenciación, que como se ve en la figura 1[b], alcanza un valor máximo del 13.70% lo cual genera un decaimiento en la calidad de los datos de las 40 pb en adelante. Para realizar el proceso de filtrado de los datos, y mejorar la calidad de los mismos se procedió a usar el software Timomatic.Como se muestra en la figura 1[c] y 1[d], se logró mejorar la calidad de los datos sustancialmente, ya que ahora todos se encuentran en el percentil superior esto se debe al corte las bases al final de una lectura, que están por debajo de un umbral de calidad que para este caso se tomó el puntaje de phred de 30. Esta herramienta permite el corte de adaptadores y otras secuencias específicas de illumina; el contenido de adaptadores

disminuyó en un 2,5% ya que para este caso el valor máximo que se encontró fue de 11,2%. La secuenciación produce una colección de secuencias sin contexto genómico, es decir no se sabe a qué parte del genoma corresponden las secuencias, por eso el mapear las lecturas de un experimento a un genoma de referencia es un paso muy importante en el análisis de datos genómicos[30]. Las lecturas del extremo emparejado se mapean utilizando el software Bowtie2 . Después de realizar el ensamble, se encontró que el porcentaje de las lecturas que se mapean para ambas muestras, SinAc y Con Ac fueron 94.16% y 92.62%, respectivamente.

6.2 Identificación de genes sobreexpresados diferencialmente.

Se encontró una lista de genes sobreexpresados, usando la herramienta EdgeR. El número de genes desregulados encontrados, (sobre-regulados/sub-regulados) fue de 1505 genes. La hipótesis que se maneja es que los genes involucrados en la biosíntesis de AC van a tener niveles más altos de expresión. Se usó el valor de logFC calculado por la herramienta EdgeR para poder encontrar dichos genes. Un total de 587 genes fueron sobre regulados en las condiciones favorables del cultivo de soja y 918 genes fueron sub-regulados. Se sabe que si los genes están funcionalmente relacionados o involucrados en la misma vía o controlados por el mismo programa regulador transcripcional, se activan o desactivan al mismo tiempo (Weirauch, 2011). Varios genes que se conocen o pueden estar involucrados directa o indirectamente regulando la producción de CA se encuentran sobre expresados (Tabla 1). Para las primeras etapas de formación de AC se encontró una sobreexpresión del gen que codifica para la Ácido proclavamínico amidinohidrolasa (*pah2*), y de el gen *bls* que codifica para la Carboxietil-arginina β-lactama-sintasa, que es la encargada de el anillo monocíclico de β-lactama además de estar relacionado con los procesos biosíntesis celular [37]. también Se observó que el regulador, *claR* (SCLAV_4181), junto con la N-glycyl-clavaminic acid sintetasa *gcas* (SCLAV_4181) y clavulanate-9-aldehyde reductase *car* (SCLAV_4190), estaban sobre-regulados, estos dos últimos genes, son muy importantes ya que se encuentran directamente involucrados en el último paso de la vía biosintética de CA. Del mismo modo, el regulador global *ccaR* (SCLAV_4204) se reguló positivamente para la condición de cultivo de soja.

Tabla 1.Se presenta, el nombre de los 1 transcritos más diferencialmente expresados que hacen parte del proceso de síntesis de ácido clavulánico. ordenados en función del valor obtenido para cada prueba realizada en dichos transcritos. La columna FDR proporciona la probabilidad de error, la cual debe ser inferior a 0.05. Por otro lado, la columna logFC, correspondiente al PdD-fold-change, muestra el cambio en la proporción de lecturas para ambas condiciones en función del PdD. Los transcritos para los cuales el valor logFC es negativo, son aquellos transcritos que se expresaron en las condiciones de cultivo con limitación de nutrientes; ocurriendo todo lo contrario para aquellos transcritos cuyo valor logFC es positivo, pues determina que dichos transcritos se expresan más en medio de cultivo de soja.

Gene ID	**Protein Name**	**log FC**	***p*-Value**	**FDR**

Up-regulated genes

Biosintesis de Acido Clavulanico.

SCLAV_4194	Clavaminate synthase 2	4.276	2.266e-10	6.496e-09
SCLAV_4181	biotin carboxylase	3.918	5.739e-11	2.087e-09
SCLAV_4187	Beta-lactamase	3.706	6.411e-11	2.286e-09
SCLAV_4189	Cytochrome P450-SU2	3.405	1.375e-08	1.583e-07
SCLAV_4191	Transcriptional activator ClaR	3.403	1.423e-08	1.631e-07
SCLAV_4182	DUF482 domain-containing protein	3.383	1.289e-08	1.494e-07
SCLAV_4196	Carboxyethyl-arginine beta-lactam-synthase	3.287	5.549e-09	7.116e-08
SCLAV_4192	Transcriptional activator ClaR	3.205	2.387e-09	3.343e-08
SCLAV_4195	Proclavaminate amidinohydrolase	3.129	1.389e-09	2.04e-08
SCLAV_4186	integral membrane protein DUF6	3.021	3.931e-11	1.504e-09
SCLAV_4190	Clavaldehyde dehydrogenase	2.805	2.228e-15	5.401e-13
SCLAV_4197	Carboxyethylarginine synthase	2.530	3.443e-10	8.069e-09
SCLAV_4193	Glutamate N-acetyltransferase 2 beta chain	1.845	6.681e-09	8.262e-08
SCLAV_4183	DUF482 domain-containing protein	1.718	7.977e-07	7.3440e-06

Down-regulated genes

SCLAV_2887	Penicillin-binding protein	-5.575	9.63e-15	5.75e-07
SCLAV_1719	Phosphate starvation-induced protein	-4.788	9.21e-06	4.87e-07
SCLAV_0091	cytochrome P450	-3.98	7.90e-15	2.10e-05
SCLAV_p1074	Carboxyethyl arginine synthase isoenzyme 1	-2.93	5.31e-10	1.88e-05
SCLAV_2155	Acyl-CoA dehydrogenase	-2.932	5.18e-07	5.43e-13
SCLAV_4082	RNA polymerase sigma factor RpoE	-2.926	5.10e-10	5.40e-13
SCLAV_2324	regulatory protein, FmdB family	-2.851	4.96e-12	1.77e-08
SCLAV_4845	Acetyl-CoA acetyltransferase	-2.822	4.95e-10	4.44e-07

En cuanto a la producción de arginina el cual es uno de los precursores para la producción de AC se encuentro una fuerte expresión de (*argj*) que cataliza dos actividades que están involucradas en la versión cíclica de la biosíntesis de arginina: la síntesis de N-acetilglutamato a partir de glutamato y acetil-CoA como donante de acetilo, y de ornitina por transacetilación entre N (2) -acetilloritina y glutamato. Así mismo la N-acetil-L-glutamato 5-fosfotransferasa (*argB*) ; que cataliza la fosforilación dependiente de ATP de N-acetil-L-glutamato.

6.3 Red de Co-expresión de genes y Análisis de enriquecimiento.

Una comprensión global de la función celular requiere el conocimiento de todas las interacciones funcionales entre las proteínas expresadas. En una red de co-expresión, los genes están unidos si sus niveles de expresión están correlacionados. Se construyó una red de correlación, usando La base de datos STRING ya que esta tiene como objetivo recopilar e integrar esta información, consolidando datos de asociación proteína-proteína conocidos, las asociaciones en STRING incluyen interacciones directas (físicas), así como interacciones indirectas (funcionales), siempre que ambas sean específicas y biológicamente significativas[38].

Se construyó la red de interacción para los genes sobreexpresados, usando la lista de genes diferencialmente expresados figura 3. La red cuenta con 524 nodos, los cuales representan los genes que se encuentran en el análisis de expresión diferencial y tienen un puntaje mayor a 0.7, esto quiere decir que solo se generó la red con aquellos genes que tenían una fuerte interacción entre sí. Gracias a esto se pudieron acoplar los resultados estadísticos, ya que la red resultante, se genero en función de los valores generados en la tabla 1. Esta red representa a aquellos genes que se sobreexpresan más que otros, basándonos en el logFC, para una mejor visualización de los datos aquellos genes con un valor de logFC alto muestran un tono más oscuro que aquellos que no tuvieron un nivel de

sobreexpresión tan elevado (amariillo). El total de genes que se conocen para la biosíntesis de AC lograron conectarse con éxito en la red, además de aquellos que están involucrados en los procesos de crecimiento celular y producción de otros metabolitos secundarios como lo es el caso de los genes involucrados en la producción de cefamicina c figura 4.
Además de aquellos nodos que se encuentran interconectados entre sí la red también muestra un conjunto de genes aislados es decir que no tienen conexiones o bordes. La aparición de nodos isla aquí fortalece el hecho de que, aunque estos nodos / genes se sobreexpresan en todos los conjuntos de datos, su correlación no pudo detectarse en función del umbral considerado.

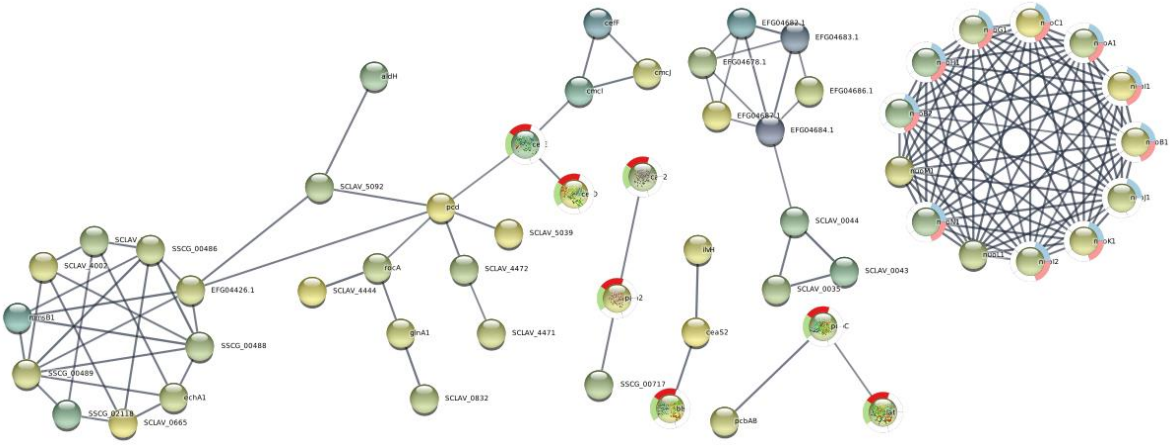

figura 3 Red de interacción de los genes más importantes, obtenidos usando StringApp, como se ve se conserva el formato de la base de datos de string, además de esto se usó un valor de confidencialidad de 0.7, lo cual nos permitió acotar la red.

Se observó que el regulador CA específico de la ruta, claR (SCLAV_4191), junto con los genes de codificación tardía ácido N-glicil-clavamínico sintetasa *gcas* (SCLAV_4181) y clavulanate-9-aldehyde reductase *car* (SCLAV_4190), estaban regulados; los dos últimos están involucrados en el último paso de la ruta biosintética de CA. Del mismo modo, el regulador global ccaR (SCLAV_4204) se reguló positivamente para la condiciones favorables del medio de cultivo. También se encontró la expresión diferencial de *argJ* que cataliza para la proteína ArgJ que es un Proteína bifuncional de biosíntesis de arginina; ya que es la encargada de catalizar Cataliza dos actividades que están involucradas en la versión cíclica de la biosíntesis de arginina: la síntesis de N-acetilglutamato a partir de glutamato y acetil-CoA como donante de acetilo, y de ornitina por trans-acetilación entre N (2) -acetilloritina y glutamato[30]. Otro precursor importante para la biosíntesis de la arginina que se encontró fue *assY* que codifica para la Citrulina-aspartato ligasa; que ayuda al transporte de aminoácidos y hace parte del metabolismo de arginina y prolina junto con SCLAV_2388.

Para agrupar las proteínas en la red en función de sus interacciones desde STRING, se usó ClusterMaker2 donde se ejecutó el agrupamiento de Markov (MCL).Se aumentó el valor de inflación a 4.0 para reducir el tamaño del clúster, establecimos fuentes de matriz para usar el atributo de puntaje de confianza STRING como ponderaciones, verificamos la opción para crear una nueva red en clúster y dejamos todas las demás configuraciones por defecto. La red resultante se simplifica enormemente y es mucho más fácil de visualizar, ya que solo se retienen las interacciones 523 nodos dentro de los clústeres. (Fig. 4).

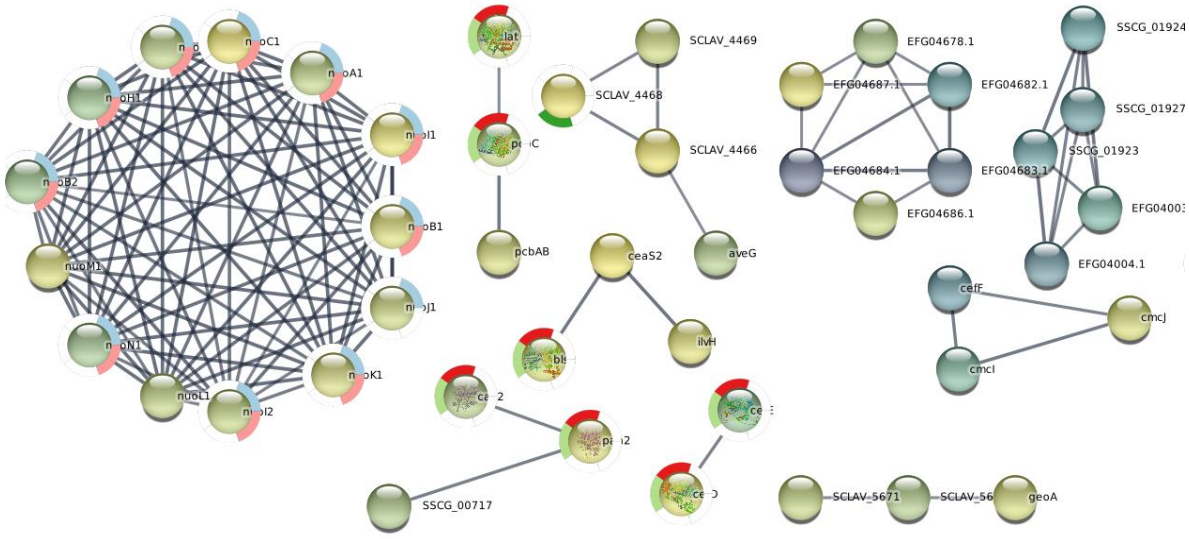

figura 4. Red de interacción de los genes más importantes, obtenidos usando StringApp, como se ve se conserva el formato de la base de datos de string, además se agruparon los genes más importantes usando la herramienta cluster Marker2.

6.4 Organización de Grupos de Genes.

El alto coeficiente de agrupación junto con una longitud los grupos formados, indica que la red generada para la biosíntesis de AC muestra una arquitectura de mundo pequeño (las redes de mundo pequeño son redes que tienen un alto coeficiente de agrupamiento, una longitud de camino característica corta y una distribución de grados de nodo siguiendo la distribución poissoniana (Albert y Barabási, 2002 ; Liao et al., 2017). Los nodos que forman una red de mundo pequeño están asociados con importantes consecuencias biológicas, ya que este nivel de conectividad permite un flujo eficiente y rápido de señales dentro de la red. Este es el caso de (SCLAV_4468, SCLAV_4466, SCLAV_4469) que son un grupo de genes fuertemente correlacionados con Tioesterasa *aveG* la cual está implicada en la biosíntesis de metabolitos secundarios, el transporte y catabolismo. Otro grupo que se encontró fue el de la Deacetoxicefalosporina C sintasa *(cefE)* ; la cual cataliza el paso entre la penicillina N y deacetoxy- cephalosporin C; con la Isopenicilina N epimerasa *(cefD)* que Cataliza la isomerización reversible entre isopenicilina N y penicilina N; que pertenece a la familia de aminotransferasa.

Otro descubrimiento importante fue la sobreexpresión del regulador transcripcional *(ccaR)* el cual se conoce que se encarga de corregular tanto la expresión de AC como de cefamicina C; esto se debe a que el grupo de genes responsables de la biosíntesis del ácido clavulánico se encuentra inmediatamente adyacente al grupo de genes de cefamicina C en el cromosoma de S. clavuligerus[39]. El gen *ccaR* del grupo de genes de cefamicina C codifica a un miembro de la familia de la proteína reguladora de antibióticos Streptomyces (SARP) y es necesario para la producción de cefamicina C y CA. Además de esto, es el encargado de regular la producción de CA tanto directa como indirectamente controlando la expresión de los genes "tempranos" del grupo de genes de CA.

La asociación de los genes juntos en los grupos de coexpresión a su vez da una indicación positiva de los procesos biológicos y las vías en las que los genes no caracterizados podrían estar involucrados al allanar el camino para que los biólogos experimentales planifiquen sus estudios.

Conclusiones.

Gracias a este análisis, se puede llegar a tener un entendimiento holístico del proceso de biosíntesis de AC, y nos permite entender como todo el sistema está interconectado entre sí y cómo se complementan todos los genes, formando clusters e interactuando con los demás, esto gracias a que no se centra en un gen en específico.

Se encontró la sobreexpresión del regulador transcripcional *(ccaR)* el cual se conoce que se encarga de corregular tanto la expresión de AC como de cefamicina C; esto se debe a que el grupo de genes responsables de la biosíntesis del ácido clavulánico se encuentra inmediatamente adyacente al grupo de genes de cefamicina C en el cromosoma de S. clavuligerus. Esto es importante ya que se en futuros análisis se podría considerar la sobreexpresión de dicho gen, o la supresión del mismo para analizar así como cambia la producción de AC y cefamicina C, así como los demás genes involucrados en las etapas tardías de producción de AC.

Los cluster son agrupaciones de genes, que comparten una función generalizada, por el cual los y pueden coordinar la expresión de genes proximales. Comprender qué genes se co-transcriben como parte de un solo mensaje proporciona información sobre la relación funcional y la regulación de los genes., es por esto que el análisis de red, y diversos métodos computacionales han demostrado ser útiles como herramientas eficientes para predecir este tipo de interacciones.

En base a estos resultados, parece que los genes que codifican las enzimas tempranas de la ruta biosintética, la parte de la ruta que es común tanto al ácido clavulánico como a los metabolitos de clavam, tienen parálogos como lo es el caso de cas2[3]. De forma contraria, los genes que codifican las enzimas biosintéticas específicas para los pasos que convierten el ácido clavamínico en ácido clavulánico son únicos. Esto se puede dar, debido a que *Streptomyces clavuligerus* contiene dos grupos de genes biosintéticos, uno para la producción de ácido clavulánico y otro para la producción de los otros metabolitos de clavam. En este sentido, el estudio de los perfiles de expresión génica o transcriptoma contenidos en el genoma de *S. clavuligerus*, podría contribuir a la identificación de genes claves expresados durante proceso celular determinado, además de contribuir a los avances en ingeniería metabólica y la mejora de cepas para aumentar la producción de ácido clavulánico.

Bibliografía.

1. Reading, C., & Cole, M. (1977). Clavulanic acid: a beta-lactamase-inhibiting beta-lactam from Streptomyces clavuligerus. Antimicrobial agents and chemotherapy, 11(5), 852- 857.

2. Vivancos, A. P., Güell, M., Dohm, J. C., Serrano, L., & Himmelbauer, H. (2010). Strand-specific deep sequencing of the transcriptome. Genome research, 20(7), 989-999.

3. Pinilla, L., Toro, L. F., Laing, E., Alzate, J. F., & Ríos-Estepa, R. (2019). Comparative Transcriptome Analysis of Streptomyces Clavuligerus in Response to Favorable and Restrictive Nutritional Conditions. Antibiotics, 8(3), 96.

4. Saudagar P., Survase S., and Singhal R. "Clavulanic acid: a review.," Biotechnol. Adv., vol. 26, no. 4; 335–51, 2008.

5. Li R., Khaleeli N., and Townsend C. "Expansion of the clavulanic acid gene cluster: identification and in vivo functional analysis of three new genes required for biosynthesis of clavulanic acid by S. clavuligerus." J. Bacteriol., vol. 182; 14: 4087–4095, 2000.

6. Wang, Z.; Gerstein, M.; Snyder, M. RNA-Seq: A revolutionary tool for transcriptomics. Nat. Rev. Genet. 2009, 10, 57–63.

7. edema, M.H.; Alam, M.T.; Heijne, W.H.M.; van den Berg, M.A.; Müller, U.; Trefzer, A.; Bovenberg, R.A.L.; Breitling, R.; Takano, E. Genome-wide gene expression changes in an industrial clavulanic acid overproduction strain of Streptomyces clavuligerus. Microb. Biotechnol. 2011, 4, 300–305.

8. J. Huang, J. Wang, Y. Li, Y. Kang, Z. Liu Transcriptomic responses to heat stress in rainbow trout Oncorhynchus mykiss head kidney head kidney Fish Shellfish Immunol., 82 (2018), pp. 32-40

9. Kibinge, N., Ono, N., Horie, M., Sato, T., Sugiura, T., Altaf-Ul-Amin, M., ... & Kanaya, S. (2016). Integrated pathway-based transcription regulation network mining and visualization based on gene expression profiles. Journal of biomedical informatics, 61, 194-202.

10. R. Kuner, T. Muley, M. Meister, M. Ruschhaupt, A. Buness, E.C. Xu, P. Schnabel, A. Warth, A. Poustka, H. Sültmann, et al. Global gene expression analysis reveals specific patterns of cell junctions in non-small cell lung cancer subtypes Lung Cancer, 63 (1) (2009), pp. 32-38

11. J. Ruan, A.K. Dean, W.X. Zhang A general co-expression network-basedapproach to gene expression analysis: comparison and applications BMC Syst. Biol., 4 (2010), p. 8

12. Liras C., Gomez J., and Santamarta I. "Regulatory mechanisms controlling antibiotic production in S. clavuligerus." J. Ind. Microbiol. Biotechnol., vol. 35, no. 7; 667–76, 2008.

13. A. Paradkar, "Clavulanic acid production by S. clavuligerus: biogenesis, regulation and strain improvement.," J. Antibiot.1–10, 2013.

14. Song J., Jensen S., and Lee K. "Clavulanic acid biosynthesis and genetic manipulation for its overproduction.," Appl. Microbiol. Biotechnol., vol. 88, 3: 659–569, 2010.

15. Caicedo-Montoya, C., Pinilla, L., Toro, L. F., Yepes-García, J., & Ríos-Estepa, R. (2019). Comparative Analysis of Strategies for De Novo Transcriptome Assembly in Prokaryotes: Streptomyces clavuligerus as a Case Study. High-Throughput, 8(4), 20.

16. Haas, B.J.; Papanicolaou, A.; Yassour, M.; Grabherr, M.; Blood, P.D.; Bowden, J.; Couger, M.B.; Eccles, D.; Li, B.; Lieber, M.; et al. De novo transcript sequence reconstruction

from RNA-Seq: Reference generation and analysis with Trinity. Nat. Protoc. 2013, 8, 1–43.

17. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nature Reviews Genetics. 2009;10(1):57–63. pmid:19015660

18. Costa-Silva, J., Domingues, D., & Lopes, F. M. (2017). RNA-Seq differential expression analysis: An extended review and a software tool. PloS one, 12(12), e0190152.

19. Koch, C. M., Chiu, S. F., Akbarpour, M., Bharat, A., Ridge, K. M., Bartom, E. T., & Winter, D. R. (2018). A beginner's guide to analysis of RNA sequencing data. American journal of respiratory cell and molecular biology, 59(2), 145-157.

20. Huang, Z., Ma, A., Yang, S., Liu, X., Zhao, T., Zhang, J., ... & Xu, R. (2020). Transcriptome analysis and weighted gene co-expression network reveals potential genes responses to heat stress in turbot Scophthalmus maximus. Comparative Biochemistry and Physiology Part D: Genomics and Proteomics, 33, 100632.

21. P. Langfelder, S. Horvath WGCNA: an R package for weighted correlation network analysis BMC bioinformatics., 9 (2008), p. 559

22. R. Pérez-Redondo, A. Rodríguez-García, J.F. Martín, P. Liras The claR gene of Streptomyces clavuligerus, encoding a LysR-type regulatory protein controlling clavulanic acid biosynthesis, is linked to the clavulanate-9-aldehyde reductase (car) gene Gene, 211 (1998), pp. 311-321

23. M.T. López-García, I. Santamarta, P. Liras Morphological differentiation and clavulanic acid formation are affected in a Streptomyces clavuligerus adpA deleted mutant Microbiology, 156 (2010), pp. 2354-2365.

24. F.J. Pérez-Llarena, P. Liras, A. Rodríguez-García, J.F. Martín A regulatory gene (ccaR) required for cephamycin and clavulanic acid production in Streptomyces clavuligerus: amplification results in overproduction of both β-lactam compounds J Bacteriol, 179 (1997), pp. 2053-2059.

25. N.L. Ferguson, L. Peña-Castillo, M.A. Moore, D.R.D. Bignell, K. Tahlan Proteomics analysis of global regulatory cascades involved in clavulanic acid production and morphological development in Streptomyces clavuligerus J Ind Microbiol Biotechnol, 43 (2016), pp. 537-555, 10.1007/s10295-016-1733-y

26. Jnawali, H. N., Liou, K., & Sohng, J. K. (2011). Role of σ-factor (orf21) in clavulanic acid production in Streptomyces clavuligerus NRRL3585. Microbiological research, 166(5), 369-379.

27. Ünsaldı, E., Kurt-Kızıldoğan, A., Voigt, B., Becher, D., & Özcengiz, G. (2017). Proteome-wide alterations in an industrial clavulanic acid producing strain of Streptomyces clavuligerus. Synthetic and systems biotechnology, 2(1), 39-48.

28.Bérénice Batut, 2020 Quality Control (Galaxy Training Materials). /training-material/topics/sequence-analysis/tutorials/quality-control/tutorial.html Online; accessed Fri Jun 05 2020

29.Batut et al., 2018 Community-Driven Data Analysis Training for Biology Cell Systems 10.1016/j.cels.2018.05.012

30.Joachim Wolff, Bérénice Batut, Helena Rasche, 2020 Mapping (Galaxy Training Materials). /training-material/topics/sequence-analysis/tutorials/mapping/tutorial.html Online; accessed Sat May 16 2020

31.Bolger, A.M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina Sequence Data. Bioinformatics, btu170.

32.Medema, M.H.; Trefzer, A.; Kovalchuk, A.; Van Den Berg, M.; Müller, U.; Heijne, W.; Wu, L.; Alam, M.T.; Ronning, C.M.; Nierman, W.C.; et al. The sequence of a 1.8-Mb bacterial linear plasmid reveals a rich evolutionary reservoir of secondary metabolic pathways. Genome Biol. Evol. 2010, 2, 212–224. [Google Scholar] [CrossRef]

33.Anders, S., Pyl, P. T., & Huber, W. (2015). HTSeq—a Python framework to work with high-throughput sequencing data. Bioinformatics, 31(2), 166-169.

34.Mark D. Robinson, Davis J. McCarthy, Gordon K. Smyth, edgeR: a Bioconductor package for differential expression analysis of digital gene expression data, Bioinformatics, Volume 26, Issue 1, 1 January 2010, Pages 139–140, https://doi.org/10.1093/bioinformatics/btp616

35.Sánchez Santana, S.d.C. (2015). Análisis de datos de RNA-Seq comparación de métodos para el estudio de expresión génica diferencial. (Trabajo Fin de Grado Inédito). Universidad de Sevilla, Sevilla.https://idus.us.es/bitstream/handle/11441/40809/S%c3%a1nchez%20Santana%20Sara%20del%20Carmen%20TFG.pdf?sequence=1&isAllowed=y

36.Langmead, Ben and Salzberg, Steven L (2012). Fast gapped-read alignment with Bowtie 2. In Nature Methods, 9 (4), pp. 357â359. [doi:10.1038/nmeth.1923][Link]

37.Jensen, S. E., Elder, K. J., Aidoo, K. A., & Paradkar, A. S. (2000). Enzymes Catalyzing the Early Steps of Clavulanic Acid Biosynthesis Are Encoded by Two Sets of Paralogous Genes inStreptomyces clavuligerus. Antimicrobial agents and chemotherapy, 44(3), 720-726.

38.R. Albert, A.L. Barabási Statistical mechanics of complex networks Rev. Mod. Phys., 74 (1) (2002), p. 47

39.Szklarczyk, D., Morris, J. H., Cook, H., Kuhn, M., Wyder, S., Simonovic, M., ... & Jensen, L. J. (2016). The STRING database in 2017: quality-controlled protein–protein association networks, made broadly accessible. Nucleic acids research, gkw937.

40.X. Liao, A.V. Vasilakos, Y. He Small-world human brain networks: perspectives and challenges Neurosci. Biobehav. Rev., 77 (2017), pp. 286-300