UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

FACULTAD DE ODONTOLOGÍA

VARIABLES ASOCIADAS A LA CALIDAD DE DATOS DE SECUENCIACIÓN EN ESTUDIOS DE METATAXONÓMICA.

T E S I S

QUE PARA OBTENER EL TÍTULO DE

C I R U J A N A D E N T I S T A

P R E S E N T A:

LUZ KARINA VELASCO CALDERÓN

TUTOR: Dra. GABRIELA ELISA MERCADO CELIS

ASESOR: Mtro. EZEQUIEL ALEJANDRO PÉREZ IBARRA

MÉXICO, Cd. Mx. **2022**

Agradecimientos

A mi tutora Gabriela Mercado Celis por la confianza y el apoyo durante el desarrollo de este proyecto, por darme la oportunidad de ser parte del Laboratorio de Genómica Clínica y compartir sus enseñanzas y conocimiento conmigo. Mi admiración hacia usted.

A Ezequiel Ibarra por la confianza y el apoyo incondicional, además de una admirable disposición a enseñar y compartir su conocimiento para poder desarrollar este proyecto exitosamente, siempre guiándome pacientemente y motivándome. Muchas gracias.

Al Laboratorio de Genómica Clínica por permitirme desarrollar este proyecto y retroalimentar cada paso durante los seminarios.

A la Universidad Nacional Autónoma de México por mi formación académica.

Agradecimientos Personales

A Dios, porque sin él nada es posible, por ser el pilar de mi vida y acompañarme en cada etapa de la misma, gracias por no soltarme en ningún momento, para él es todo el honor.

A mis padres Román Velasco Martinez y María de Jesús Calderón Villegas por brindarme su amor y apoyo incondicional, por el esfuerzo de sostener mis estudios, apostando lo mejor para mí. Ahora esas horas esperándome fuera de la facultad se materializan en metas cumplidas, sin ustedes nada sería posible y cada logró es para y por ustedes, eternamente agradecida por ser su más grande apuesta al igual que mis hermanos, por el optimismo y creer en mí más de lo que yo lo puedo hacer.

A mis hermanos, José Carlos Velasco Calderón por guiarme y procurarme, por tomar mis retos como suyos y no dejarme caer ni conformarme. Por ser un gran ejemplo de perseverancia, resiliencia y hacerme saber que soy capaz de lograr lo que me proponga y a Luis Román Velasco Calderón, gracias por ser mis compañeros de vida y brindarme su apoyo incondicional.

A Teresa Velasco Martínez, por el apoyo que me ha brindado en cada etapa de mi vida, el cariño incondicional, procurarme y tenerme presente pese a la distancia, gracias por siempre estar.

A Domingo Calderón Avilés, porque en cada recuerdo estás presente y en cada consejo vives en mí. Eres inefable y tengo la certeza de lo orgulloso que te sientes desde donde estés.

A Jennifer Osorio Ortega por motivarme a seguir en todo momento, el apoyo incondicional, ser un respiro en los malos momentos y la mejor compañía para celebrar los buenos.

A Mariana Bernabé Rodríguez por alegrarme la vida, alentarme en todo momento y confiar en mí, demostrándome que no es cuestión de tiempo si no de incondicionalidad.

ÍNDICE

Índice de Ilustraciones

Índice de Tablas

Glosario

- Amplicón

Fragmento de ADN amplificado por la reacción en cadena de la polimerasa (PCR) o cualquier otro proceso que dé lugar a la producción de diferentes copias de ese fragmento.

- Bray Curtis

La diferencia total en la abundancia de especies entre dos sitios, dividido para la abundancia total en cada sitio.

- Cebador aleatorio

Los cebadores aleatorios consisten en una mezcla de oligonucleótidos que representan todas las posibles secuencias hexámeras. Los cebadores hexámeros aleatorios se usan comúnmente para cebar DNA o RNA monocatenario para la extensión por DNA polimerasas o transcriptasas inversas.

- Ciencias Omicas

Son las ciencias que permiten estudiar un gran número de moléculas, implicadas en el funcionamiento de un organismo, como genes, proteínas y metabolitos, permitiendo la creación de la genómica, proteómica, metabolómica, entre otras.

- Diversidad Alfa

Es la diversidad media de especies en un sitio a escala local.

- Diversidad Beta

Se aplica al grado de cambio en la composición de una comunidad entre una unidad de muestreo y otra a lo largo de gradientes, o a la variación en la composición de la comunidad entre las unidades de muestreo.

- Enzimas de restricción

Una enzima de restricción es una proteína aislada a partir de bacterias que cortan secuencias de ADN en sitios específicos de la secuencia, lo que produce fragmentos de ADN con una secuencia conocida en cada extremo.

- Genoma

Secuencia de nucleótidos que constituye el ADN de un individuo o de una especie

- Ilumina

Desarrolla, fabrica y comercializa sistemas integrados para el análisis de la variación genética y la función biológica.

- Índice de Shannon

Se usa en ecología u otras ciencias similares para medir la biodiversidad específica, que en la mayoría de los ecosistemas naturales varía entre 0,5 y 5.

- Índice de Simpson

Es un índice de dominancia más que de diversidad y representa la probabilidad de que dos individuos escogidos al azar pertenezcan a la misma especie.

- Material Genético

Cualquier material de origen vegetal, animal o microbiano u otro que tenga información genética y que la transmita de una generación a la siguiente. Esa información controla la reproducción, el desarrollo, el comportamiento, etc.

- *Phred Quality Score*

El nivel de calidad Phred es una medida de calidad en la identificación de las nucleobases generadas por la secuenciación automatizada de ADN.

- Regiones hipervariables V3 y V4

Zonas especificas de variabilidad genética del ARN ribosomal 16S. Siendo las regiones V3 y V4 las que presentan mayor variabilidad, permitiendo obtener una asignación taxonómica mas precisa.

- Software RStudio

RStudio es un entorno de desarrollo integrado para el lenguaje de programación R, diseñado para hacer análisis estadísticos y gráficos.

- *Taq* Polimerasa

La *Taq* polimerasa es como se conoce comúnmente a la enzima ADN polimerasa de *Thermus aquaticus*, una bacteria. Es la polimerasa más frecuente y de bajo costo que se puede usar para realizar una PCR.

- Taxa

Es el plural latino de taxón, el término es usado en la terminología de la clasificación biológica para referirse a un grupo de organismos de cualquier rango.

- TM7x

Nanosynbacter lyticus tipo cepa TM7x. Es un filotipo de uno de los filos más enigmáticos de *Candidatus Saccharibacteria*, es el único miembro del filo que se ha cultivado con éxito a partir de la cavidad oral humana y se ha mantenido estable *in vitro.*

- Unifrac ponderado

UniFrac es una métrica de distancia utilizada para comparar comunidades biológicas, cuantitativamente, donde toma en cuenta la abundancia de organismos encontrados.

- Unifrac no ponderado

UniFrac es una métrica de distancia utilizada para comparar comunidades biológicas, cualitativamente solo considerando la presencia o ausencia de microorganismos.

Glosario de abreviaturas

- ASV

“Variantes de Secuencia de Amplicón” se refiere a una secuencia única que se asigna a un grupo taxonómico. Los métodos ASV infieren las secuencias biológicas en la muestra antes de la introducción de los errores de amplificación y secuenciación, distinguiendo las variantes de secuencia que difieren en tan solo un nucleótido.

- *FAIR*

Por sus siglas en inglés “Principios de localización, accesibilidad, interoperabilidad y reutilización”.

- IVS

Por sus siglas en inglés “Secuencias de intervención, decir un intrón” (región en el interior de un gen).

- 16S rARN

El ARN ribosomal 16S (ARNr 16S o 16S rRNA) es el componente de la subunidad menor (30S). La secuenciación del ADNr 16S es particularmente importante en el caso de bacterias con perfiles fenotípicos inusuales, bacterias raras, bacterias de crecimiento lento, bacterias no cultivables e infecciones con cultivos negativos.

- rcADN

El ADN circular puede encontrarse en forma relajada o en forma superenrollada. En la forma relajada, el círculo se halla desplegado sobre un único plano; en la forma superenrollada el contorno del círculo va girando sobre sí mismo de manera tal que adquiere profundidad.

- rADN

El ADN recombinante (rADN) es una tecnología que utiliza enzimas para cortar y unir secuencias de ADN de interés. Las secuencias de ADN recombinado se pueden

colocar en unos vehículos llamados vectores que transportan el ADN hacia el lugar adecuado de la célula huésped donde puede ser copiado o expresado.

- RNasas

Las RNasas son proteínas con actividad enzimática presentes en bacterias, hongos, plantas superiores y mamíferos, que participan en procesos fisiológicos diversos tales como: muerte celular, replicación del DNA, transcripción, procesamiento y edición del RNA, defensa del y control del crecimiento tumoral.

- rrn

Operones de ARN ribosomal.

- RT-PCR

Método que se usa para hacer muchas copias de una secuencia genética específica con el fin de analizarla. Se usa una enzima llamada retrotranscriptasa que convierte un trozo específico de ARN en un trozo de ADN compatible. Luego, otra enzima llamada ADN–polimerasa amplifica ese trozo de ADN. Las copias de ADN amplificadas ayudan a identificar si hay un gen que produce la molécula específica de ARNm.

- PCoA

Análisis de coordenadas principales

- PCR

Son las siglas por las que se conoce a la reacción en cadena de la polimerasa. Esta técnica permite amplificar pequeñas regiones específicas del ADN en laboratorio. Es decir, consigue que un pequeño segmento de ADN que pasaría desapercibido en un análisis cualquiera se multiplique millones de veces y así sea fácil de detectar.

- OTU

“Unidad Taxonómica Operacional” es el grupo de organismos actualmente en estudio. En este sentido, una OTU es una definición pragmática para agrupar individuos por similitud.

- Valor de “p”

Si el valor p es menor que 0,05, rechazamos la hipótesis nula de que no hay diferencia entre las medias y concluimos que sí existe una diferencia significativa. Si el valor p es mayor que 0,05, no podemos concluir que existe una diferencia significativa.

- VPH

Virus del Papiloma Humano.

1.O Resumen

Los estudios de Metataxonómica son procesos de alto rendimiento los cuales nos permiten caracterizar la microbiota, creando árboles metataxónomicos y mostrando las relaciones entre todas las secuencias obtenidas. Son una opción rápida y económica cuando la microbiota es conocida, además de tener el potencial de mejorar los diagnósticos, al ser un procedimiento relativamente nuevo aún no se cuentan con protocolos estandarizados sobre los pasos que conllevan, por lo cual algunos de estos podrían introducir sesgos o inhabilitar su reproducibilidad.

Por lo tanto, se decidió analizar la calidad en las lecturas secuenciadas del gen 16S ARNr por secuenciación masiva paralela de la región hipervariable V3-V4 para la asignación taxonómica en función del método de recolección de la muestra, almacenamiento, procesamiento de muestra y de datos.

Se utilizaron 50 muestras salivales de pacientes sanos y pacientes con adenocarcinoma pulmonar del Laboratorio de Genómica Clínica, las cuales debían contar con lecturas f*orward* y *reverse,* datos de recolección, calidad de las lecturas, las cuales fueron analizadas posteriormente con el *software Rstudio*.

Se destacan aspectos importantes, como la importancia del uso de buenas prácticas al momento de la recolección de la muestra, además de la necesidad de perfeccionarlas y así mejorar la reproducibilidad de estos estudios. Así como la influencia del tiempo de almacenamiento, kit de extracción y calidad de ADN. Encontrando diferencias significativas en algunas de estas.

2. Antecedentes

2.1 Microbiota

La microbiota es un ecosistema complejo de microorganismos formado por bacterias, virus, protozoos y hongos, que viven en diferentes nichos del cuerpo humano, como el tubo gastroentérico, la piel, la boca, el sistema respiratorio y la vagina. Más del 70% de la microbiota vive en el tracto gastrointestinal en una relación mutuamente beneficiosa con su anfitrión. La microbiota juega un papel importante en muchas funciones metabólicas, incluida la modulación de la homeostasis de la glucosa y los lípidos, la regulación de la saciedad, la producción de energía y vitaminas. Ejerce un papel en la regulación de varios mecanismos bioquímicos y fisiológicos a través de la producción de metabolitos y sustancias. Además, la microbiota juega un papel importante en acciones anticancerígenas y antiinflamatorias (1-3).

Responde y se adapta a los cambios en el huésped, un ejemplo de esto es la inflamación. La alteración de la microbiota oral puede estar íntimamente asociada a enfermedades bucales y sistémicas. (1)

Los microorganismos asociados con el cuerpo humano, en específico las bacterias, se estima superan en número a las células humanas dentro de un individuo en un orden de magnitud. (2)

Microbiota oral

El microbioma oral es muy diverso e incluye bacterias, hongos, virus, arqueas y protozoos. En la cavidad bucal están presentes aproximadamente 700 especies y la mayoría de ellas son autóctonas. Entre ellas, aproximadamente el 54% han sido cultivadas y nombradas, el 14% son cultivadas, pero no nombradas, y 32% se conocen sólo como filotipos no cultivados. En la actualidad un número creciente de estudios han demostrado que la microbiota oral desempeña un papel vital en la patogénesis y desarrollo de muchas enfermedades bucales y sistémicas. (2)

Microbioma

El microbioma hace referencia a la colección de genomas de todos los microorganismos del medio. La obtención de una visión integral de los ecosistemas microbianos asociados con el cuerpo humano (el microbioma humano) se ha hecho posible gracias a los avances en los análisis “ómicos” y a las técnicas de secuenciación de nueva generación (NGS).

2.2 Métodos para el estudio del Microbioma

Metagenómica (*whole shotgun metagenomic sequencing*)

La metagenómica es la disciplina que estudia el material genético de los microorganismos que componen un ambiente particular, en un estudio metagenómico no es necesario cultivar cada organismo para poder estudiarlo. Tiene la ventaja de poder evaluar la totalidad de los organismos presentes, sin la necesidad de dirigir la búsqueda hacia un grupo particular. El análisis metagenómico cuantitativo crea un perfil de genes y especies, que permite la identificación y clasificación filogenética tanto de los conocidos como de los nuevos. **(**4)

Metataxonómica (16S rRNA gene sequencing)

Metataxonómica es un término que se propone y define como el proceso de alto rendimiento utilizado para caracterizar toda la microbiota y crear un árbol metataxonómico, que muestra las relaciones entre todas las secuencias obtenidas. Si bien los virus son una parte integral del microbiota, no hay genes marcadores virales universales disponibles para realizar tales asignaciones taxonómicas. (5) Las secuencias 16S amplificadas por PCR se agrupan en función de la similitud para generar Unidades Taxonómicas Operativas (OTU), Tablas de Variante de Amplicón (ASV) y en la reconstrucción de relaciones filogenéticas.

La metataxonómica es una opción rápida y económica cuando el microorganismo de interés es una bacteria conocida, pero si hay microorganismos desconocidos, nuevos, una mezcla de virus, hongos y bacterias, entonces la metagenómica será

una mejor opción. (6) La metataxonómica y la metagenómica, con su independencia y riqueza de datos, tienen el potencial de mejorar los diagnósticos. Pero antes de que cualquiera de estos métodos pueda integrarse por completo en los protocolos de diagnóstico, sus beneficios relativos deben compararse y validarse mediante métodos de cultivo. (6)

2.3 ¿Qué es el ARNr 16S?

El ARNr 16S es un polirribonucleótido de aproximadamente 1500 nucleótidos codificado por el gen rrs, también denominado ADN ribosomal 16S. Las secuencias 16S amplificadas por PCR se han agrupado típicamente en función de la similitud para generar unidades taxonómicas operativas (OTU) y secuencias OTU representativas en comparación con bases de datos de referencia para inferir una taxonomía probable.

Las secuencias 16S también se han explotado utilizando métodos de bajo rendimiento para distinguir cepas (a veces llamadas subespecies) basándose en polimorfismos dentro del gen. Los polimorfismos de un solo nucleótido (SNP) se han utilizado para rastrear cepas de relevancia clínica o cuando están vinculadas de manera estable a otras partes del haplotipo bacteriano, para predecir características fenotípicas, en la actualidad, la gran mayoría de los estudios secuencian solo una parte del gen, utilizando la plataforma Illumina en su mayoría.

Uso en investigación

16S (ARNr 16S), originalmente propuesto por Pace y col. (1986), fue presentado como una buena opción para la clasificación de bacterias. La idea fue rápidamente adoptada por la comunidad científica y la secuencia del ARNr 16S se ha utilizado para conformar bases de datos especializadas. Lo anterior ha permitido que las secuencias del ARNr 16S sean utilizadas como una herramienta importante en la reconstrucción de relaciones filogenéticas. Además, el uso de secuencias del ARNr 16S facilitó el establecimiento del proyecto árbol de la vida universal (*AllSpecies*

Living Tree Project), el cual se ha constituido como una referencia de relación de procariotas fácilmente organizada en bases de datos dinámicas que compilan los datos de todas las secuencias accesibles del gen ARNr 16S. (7)

Características

- Estructura secundaria que se caracteriza por tener segmentos de doble cadena que permiten la formación de asas y hélices.
- El ARNr 16S contiene nueve regiones (V1–V9) menos conservadas o hipervariables (Baker y col. 2003), que son las que aportan la mayor información útil para estudios de filogenética y taxonomía
- Las regiones conservadas son de gran ayuda para diseñar iniciadores universales que permitan la amplificación de las diversas regiones hipervariables de la gran mayoría de los ARNr 16S de los microorganismos presentes en una comunidad.
- La región del gen ARNr 16S que amplifican, tiene un efecto determinante en la descripción de la diversidad bacteriana de muestras ambientales.
- Pese a estos inconvenientes, el uso del ARNr 16S como marcador sigue siendo la herramienta más fuerte para el entendimiento de las comunidades bacterianas de todos los ambientes estudiados.
- El grado de heterogeneidad del gen de ARNr 16S intragenómico entre bacterias varía de 0 a 11,6% de divergencia de secuencia para 1 a 15 copias del gen de ARNr 16S. Se encontró que el sesenta y dos por ciento de las bacterias con más de una copia del gen del ARNr 16S mostraban algún grado de heterogeneidad intragenómica para este gen. (7)

Ventajas del 16S

- Primero, ha eliminado las variantes de secuencia de artefactos menores debido a la amplificación por PCR y los errores de secuenciación al colapsar secuencias en grupos.

- Disminuye las variantes de secuencia legítimas que existen entre taxa bacterianos estrechamente relacionados.
- Su función esencial, ubicuidad y propiedades evolutivas, le han permitido convertirse en el marcador molecular más utilizado en ecología microbiana.
- Hay múltiples copias de este gen en una bacteria determinada.
- Este procedimiento revolucionó la ecología microbiana y cambió permanentemente la forma en que estudiamos los procariotas en el medio ambiente.
- La adopción de herramientas moleculares por los ecologistas microbianos ha mejorado rápidamente nuestro conocimiento de la abundancia, diversidad y función de las procariotas.
- Los grupos taxonómicos para los que se compilaron alineaciones cubren múltiples niveles taxonómicos, incluido el dominio (Bacteria), filo (*Actinobacteria, Firmicutes y Proteobacteria*), orden (*Bacillales, Enterobacteriales y Lactobacillales*), clase (*Alphaproteobacteria* y *Gammaproteobacteria*), familia (*Chlamydiaceae* y *Mycoplasmataceae*), género (*Streptococcus*) y especies / subespecies (*Shreptococcus)*. (7)

Desventajas

- Ninguno de los métodos moleculares basados en ARNr 16S permite una representación precisa de las comunidades microbianas.
- El sesgo se introduce en el análisis de la comunidad molecular por muchos factores mecánicos, como el manejo de muestras, la fijación, la extracción de ADN y la PCR, también se crea por la existencia de múltiples copias heterogéneas del gen de ARNr 16S dentro de un genoma. (8)
- Si se elige un solo gen de ARNr 16S para representar una cepa, la filogenia obtenida puede cambiar significativamente dependiendo de qué copia se eligió.
- Dificultan la separación entre comensales estrechamente relacionados a nivel de especie, y los enfoques basados en el ARNr 16S tampoco

proporcionan información sobre rasgos bacterianos como la virulencia, la adherencia o la resistencia antimicrobiana. (7, 8)

Alternativa

Un marcador alternativo al gen de ARNr 16S para estudios de ecología microbiana molecular es el gen de copia única que codifica la subunidad β de la ARN polimerasa. (8)

2.4 Sesgos en Metataxonómica

El sesgo es una forma ubicua y particular de error sistemático que surge de las diferentes eficiencias con las que se miden varios taxa (es decir, preservados, extraídos, amplificados, secuenciados o identificados y cuantificados bioinformáticamente) en cada paso. Incluyen opciones relacionadas con la recolección de muestras, almacenamiento y preservación de muestras, extracción de ADN, cebadores amplificadores, tecnología de secuenciación, longitud y profundidad de lectura, las diferentes bases de datos para alineación de secuencias y técnicas de análisis bioinformático. (9)

Cada paso físico, químico y biológico involucrado en el análisis molecular de un ambiente es una fuente de sesgo que conducirá a una visión distorsionada del "mundo real". A continuación, se presenta un resumen de los errores reportados del enfoque ecológico molecular que pueden resultar en una descripción errónea de la diversidad de un nicho ecológico, así como las recomendaciones para limitar su efecto. (10)

Diseño de Estudios de Metataxonómica

Posibles factores de confusión en estudios en humanos: estilo de vida y factores clínicos

1. Aplicar criterios de exclusión (ejemplo, lista de criterios de exclusión del Proyecto de Microbioma Humano) (9)
2. Equilibrar los factores clínicos entre los grupos experimentales.

3. Seleccionar cuidadosamente los grupos control.

Grupo (s) control

1. El grupo control apropiado debe tener un fenotipo clínico distinto del que se está estudiando, al mismo tiempo que coincide con otros criterios relevantes para evitar factores de confusión.
2. En los estudios longitudinales, los individuos pueden usarse como su propio control mediante la recolección de muestras de referencia antes y después del tratamiento.
3. Seleccionar varios grupos control sobre varios criterios puede proporcionar más información.

Muestreo Recomendado en Estudios de Metataxonómica

La estrategia analítica necesaria para obtener una librería representativa del ecosistema depende en gran medida del número esperado de miembros de su población. El rango de diversidad microbiana se puede estimar inicialmente a partir de:

(i) El número de diferentes morfotipos determinados por microscopía de fluorescencia después de la tinción específica que se une al ADN de células vivas únicamente.

(ii) Parámetros físicos y químicos del sitio de muestreo como el pH, temperatura o sustratos suministrados.

(iii) Preselección de la diversidad mediante electroforesis en gel desnaturalizante de rADN16S amplificado. (10)

Se debe tener especial cuidado durante su transporte al laboratorio para evitar la pérdida de ácidos nucleicos debido a la lisis de las muestras.

Momento y frecuencia de la recolección de muestras

1. El muestreo transversal de pacientes es apropiado para las firmas de microbioma de diagnóstico y factible para tipos de microbioma temporalmente estables. Tiene la ventaja de que se pueden analizar más individuos.

2. Se debe elegir un muestreo repetido a lo largo del tiempo cuando se quiere obtener una idea de la dinámica temporal y una visión integral de los cambios en la comunidad microbiana. También es apropiado para monitorear la gravedad de la enfermedad o la respuesta a un tratamiento, para discriminar entre disbiosis asociada a enfermedad y estados preclínicos. Para el muestreo repetido, la frecuencia debe ser similar entre sujetos.

Procesamiento de muestras recomendado en estudios de Metataxonómica

Temperatura de almacenamiento

El almacenamiento a -80° C es un protocolo estándar, pero el impacto del almacenamiento de archivos a largo plazo, la congelación y descongelación de muestras en la integridad del microbioma sigue siendo una cuestión no resuelta. (9) La recolección y el almacenamiento de muestras (incluidos los ciclos de congelación-descongelación) tienen un impacto dependiente del tipo de muestra en la composición (diversidad y estructura) de la microbiota.

A continuación, presentamos las recomendaciones para el almacenamiento

Almacenamiento a largo plazo:

a) La norma establece que deben congelarse a −80° C

b) Otro enfoque: liofilización y almacenamiento a temperatura ambiente o −20 ° C (Evaluación de muestras regularmente) (11)

Almacenamiento a corto plazo (pocos días):

a) Almacenamiento a 4 ° C: efecto mínimo (frente a −80° C) (11)

c) Soluciones estabilizantes

d) Además de los estudios basados en ADN, los estudios que incluyen perfiles de muestras de ARN y metabolómicos pueden necesitar requisitos de almacenamiento más estrictos (por ejemplo, estabilizadores de ARN)

e) Mantener la cadena de frío: minimizar los ciclos de congelación-descongelación.

Coherencia: procedimientos, protocolos, información recopilada

1. Implementar procedimientos consistentes y minimizar las variables entre grupos.

2. Recopilar la máxima información sobre muestras y procedimientos experimentales para permitir la reproducibilidad.

Extracción de ADN

La extracción de ADN es una fuente importante de variación y sesgo. El desafío radica en la lisis eficiente de células microbianas heterogéneas en la comunidad sin dañar sus genomas (algunas bacterias pueden resistir los métodos de lisis mecánicos o químicos comunes. Se sabe que la extracción de ADN tiene un gran impacto en la abundancia aparente de algunos miembros de las comunidades microbianas, y existen muchos kits comerciales distintos. Las especies bacterianas difieren en la facilidad con que se lisan y por lo tanto, en la cantidad de ADN que producen durante la extracción de ADN y difieren en su número de copias del gen ARNr 16S y, por lo tanto, en la cantidad de producto de PCR que esperamos obtener por célula.

La lisis insuficiente o preferencial de las células puede sesgar la visión de la composición de la diversidad microbiana, ya que el ADN o el ARN, que no se libera de las células, no contribuirá al análisis final de la diversidad. Por otro lado, deben evitarse las condiciones rigurosas requeridas para la lisis celular de bacterias Gram-positivas ya que este tratamiento puede conducir a ácidos nucleicos altamente fragmentados de células Gram-negativas. Los ácidos nucleicos fragmentados son fuentes de artefactos en experimentos de transcripción inversa o amplificación por PCR y pueden contribuir a la formación de productos de PCR quiméricos. Además,

varios componentes bióticos y abióticos de los ecosistemas ambientales, como las partículas inorgánicas o la materia orgánica, afectan la eficiencia de la lisis y pueden interferir con la purificación del ADN y los pasos enzimáticos posteriores.

De acuerdo con lo anterior se proponen las siguientes recomendaciones:

1) Realizar una combinación de procedimientos de lisis químicos y mecánicos para obtener una captura de la composición comunitaria más precisa.
2) Utilizar protocolos específicos de procesamiento de muestras y kits comerciales que maximicen el rendimiento, la calidad y la integridad de la extracción de ADN de la estructura y diversidad de la comunidad microbiana.
3) Utilizar el mismo procedimiento durante todo el estudio.
4) Procesar todas las muestras:
    i) Una al lado de la otra
    ii) Utilizar el mismo lote de reactivos
    iii) Si se utilizan varios kits, trate el 'lote del kit' como un factor en el análisis estadístico
    iv) Si hay muestras de biomasa alta y baja, comprometerse con un tipo de kit para todas las muestras.

Calidad de ADN (Pureza, Integridad y Cantidad)

En particular, los sesgos en la cobertura de secuencias genómicas pueden deberse a las condiciones de crecimiento, la fase de crecimiento del organismo, el sesgo de clonación, la eficiencia de secuenciación, el número relativo de copias del genoma, la eficiencia de extracción de ADN donde se ha demostrado que la estructura de la membrana celular puede tener un efecto significativo. Además, la cantidad de ADN de un organismo que está disponible para formar parte de una biblioteca de secuenciación depende de la eficiencia del protocolo de extracción de ADN. En una muestra mixta, los organismos con paredes celulares gruesas pueden producir relativamente poco ADN, lo que lleva a una sobrerrepresentación de ese organismo en la biblioteca de secuenciación final. (12)

El método más utilizado para determinar la calidad del ADN es la relación de absorbancia a 260 y 280 nm (A260 nm/A280 nm). Los cocientes en torno a 1,8 indican que el ADN es de buena calidad, mientras que los valores más bajos indican contaminación por proteínas, ya que éstas tienen un pico de absorción a 280 nm resultante de los aminoácidos aromáticos. Los valores más altos indican contaminación de ARN. El ADN también puede evaluarse en un gel para determinar su tamaño. Como comprobación de la identidad de la muestra de ADN, se pueden determinar las repeticiones de microsatélites altamente polimórficas y compararlas con otra fuente de ADN del mismo individuo. (12)

Contaminación

La contaminación puede tener un impacto significativo al estudiar muestras con baja biomasa microbiana, como las muestras salivales. Cada kit o reactivo puede introducir contaminantes únicos, que a veces varían en su abundancia en orden de magnitud y pueden esconder la verdadera señal biológica de muestras de baja biomasa. Por lo tanto, un análisis más detallado de la contaminación de los reactivos y las contramedidas para reducirla podría ser de gran valor en esta área. (9) Sin embargo, la presencia de dicho ADN contaminante, generalmente asociado con estudios basados en PCR, no se ha informado de su posible impacto en los estudios de perfiles basados en genes de ARNr 16S de alto rendimiento. La presencia de secuencias contaminantes es mayor en muestras de baja biomasa (sangre, pulmón o saliva) que en muestras de alta biomasa (heces fecales), lo que sugiere que existe un punto crítico, en el que el ADN contaminante se vuelve dominante en las bibliotecas de secuencias. (9)

Las posibles fuentes de contaminación del ADN incluyen, agua de grado de biología molecular que esté expuesta a contaminación, reactivos de PCR y los propios kits de extracción de ADN. Por ejemplo, la contaminación del agua por géneros de bacterias del suelo incluyendo *Acinetobacter, Alcaligenes, Bacillus, Bradyrhizobium, Herbaspirillum, Legionella, Leifsonia, Mesorhizobium, Methylobacterium, Microbacterium, Novosphingobium, Pseudomonas, Ralstonia, Sphingomonas, Stenotrophomonas* y *Xanthomonas*.

Salter y cols, investigaron el impacto de la contaminación en los estudios de microbiota, así como métodos para limitarlo. Observaron que existe una variación en el contenido de contaminantes entre los laboratorios, lo que puede deberse a diferencias entre los lotes de reactivos / kits o los contaminantes introducidos por el ambiente del laboratorio. (13) Los autores proponen realizar un análisis *in silico* para los estudios de microbiota y así identificar los contaminantes que se secuenciaron utilizando controles negativos o bases de datos de contaminantes para eliminarlos del análisis final. (13) También sugieren que el conocimiento de las especies contaminantes comunes, la recopilación cuidadosa de controles para cubrir diferentes lotes de muestreo, kits de extracción y PCR además de la secuenciación para monitorear el contenido de estos controles, podría mitigar el efecto de estos contaminantes. (13)

ADN contaminante

El ADN contaminante, que contiene la secuencia diana específica de la reacción de PCR involucrada, puede conducir tanto a la amplificación en controles negativos sin que se agregue ADN externo como a la coamplificación en reacciones experimentales. El análisis directo de estos amplicones mezclados artificialmente mediante secuenciación o hibridación conduciría a resultados ambiguos, mientras que la clonación o electroforesis en gel y el análisis posterior simularían la diversidad de secuencias que en realidad no existen. (10)

El ADN no específico puede introducirse en reacciones de PCR como contaminantes de tubo a tubo, es decir, los productos de amplificación de reacciones anteriores se transfieren involuntariamente a reacciones recientes o mediante reactivos contaminados. Varios informes describieron estas últimas contaminaciones como ADN bacteriano. (10)

La amplificación de genes de ARN ribosómico parece ser extremadamente sensible a la contaminación del ADN bacteriano, ya que las regiones universalmente conservadas de genes bacterianos sirven como secuencias diana. Maiwald y col. *Taq* polimerasa es contaminada por ADN caracterizado que se amplificó durante la

PCR con un juego de cebadores para *LegionellaGen* de ARNr 5S. Sus resultados indican que no es la bacteria utilizada para la producción de la enzima recombinante sino las bacterias contaminantes del suelo fueron el origen del ADN extraño. (10)

Se han desarrollado varias estrategias para evitar o eliminar la contaminación del ADN, que incluyen tanto la organización del laboratorio como los sistemas de descontaminación, probando la confiabilidad de varios procedimientos de descontaminación donde encontraron que el tratamiento con UV y la digestión con uracilo ADN glicosilada antes de la PCR eran los más efectivos. (10)

Secuenciación

Cebadores

La amplificación selectiva (por ejemplo, 16S, 18S, ITS) de regiones génicas específicas ha sido durante mucho tiempo estándar para la secuenciación de comunidades microbianas, pero introduce sesgos y omite organismos y elementos funcionales del análisis. (11) Sinha R y cols., reportaron que la adecuada elección del cebador de amplificación 16S produjo un efecto mayor en los datos de secuenciación que en la profundidad de secuenciación, el almacenamiento y transporte de la muestra. (9)

- La elección de los cebadores depende de:
    a. Estudios previos
    b. Gen marcador al objetivo
    c. El nivel de resolución necesario para el análisis filogenético
- La elección de la región es más importante que la longitud del amplicón
- Los cebadores se dirigen a los tipos de microbios a identificar [p. Ej., 16S para bacterias, espaciador transcrito interno (ITS) para hongos]

En la amplificación por PCR del 16S rDNA / rcDNA de microbiota compleja, una mezcla de moléculas homólogas sirve como molde. El ADN amplificado sólo puede reflejar la abundancia cuantitativa de especies si las eficiencias de amplificación son las mismas para todas las moléculas. Esto requiere varios supuestos:

(i) Todas las moléculas son igualmente accesibles para la hibridación del cebador.
(ii) Los híbridos cebador-molde se forman con la misma eficacia.
(iii) La eficacia de extensión de la ADN polimerasa es la misma para todos los modelos.
(iv) Las limitaciones por el agotamiento del sustrato afectan de manera equivalente las extensiones de todas las plantillas.

Estas suposiciones parecen difíciles de mantener, ya que los cebadores universales empleados para la amplificación de rDNA / rcDNA contienen a menudo degeneraciones que pueden influir en la formación de híbridos de cebador-molde.

La eficiencia de hibridación y la especificidad de los cebadores influyen en la amplificación por PCR de las plantillas de rDNA 16S mixtas. La unión subóptima del cebador dará como resultado una amplificación menos eficaz del ADN respectivo. Especialmente los cebadores universales o específicos de dominio deben tener una eficiencia de hibridación uniforme para garantizar la amplificación de todos los ADNr 16S diana o regiones hipervariables del mismo.

PCR

Las amplificaciones por PCR pueden provocar sesgos cuando la concentración del templado de ADN es baja o el número de ciclos de PCR es alto. El número de pasos de amplificación también podría generar sesgos. Aunque es un método de rutina para cultivos puros, surgen varios problemas cuando los métodos se aplican a comunidades ambientales:

(i) Inhibición de la amplificación por PCR por contaminantes co-extraídos

(ii) Amplificación diferencial

(iii) Formación de productos de artefactos de PCR

(iv) ADN contaminante

(v) Variaciones de la secuencia del ARNr 16S debido a *rrn.* La heterogeneidad del operón conduciría inevitablemente a un reflejo sesgado de la diversidad microbiana.

Formación de artefactos de PCR

La aparición de artefactos de PCR es un riesgo potencial en el análisis de microbiota compleja mediado por PCR, ya que sugiere la existencia de organismos que en realidad no existen en la muestra investigada. Se han informado varios tipos de artefactos de PCR: (10)

(i) Quimeras entre dos moléculas homólogas diferentes.

Se ha observado ampliamente la recombinación *in vitro* de ADN homólogo que conduce a moléculas quiméricas compuestas por partes de dos secuencias diferentes, esto no se limita a la amplificación del ADNr 16S a partir de microbiota compleja.

Se pueden generar quimeras entre dos moléculas de ADN diferentes con alta similitud de secuencia (es decir, genes homólogos) durante el proceso de PCR, ya que las cadenas de ADN compiten con cebadores específicos durante la etapa de hibridación. (10)

Además de la síntesis de cadena incompleta durante el proceso de PCR, se ha sugerido que el daño del ADN promueve la formación de moléculas quiméricas en la coamplificación por PCR de moldes con similitudes de secuencia elevadas. Pääbo y col. investigaron la influencia de las roturas del molde causadas por la digestión con enzimas de restricción, la irradiación UV, la sonicación y la depurinación, sobre la coamplificación por PCR de los genes de lisozima de vaca tipo 2b y 3. Se demostró que todos los tipos de daños en el ADN respaldan la producción de productos de PCR recombinantes. Dado que es probable que las condiciones rigurosas de lisis celular para la preparación de ADN a partir de muestras ambientales causen daños similares a los descritos por Pääbo et al. Estos hallazgos podrían tener un impacto significativo en la amplificación del rDNA 16S del microbiota complejo. (10)

Concordando con el hallazgo de Liesack et al. en el que reportan la aparición de rDNA 16S quimérico en la PCR cuando se utilizó como molde DNA de baja masa molecular (4-6 kilo base) en el análisis de un cultivo mixto de dos bacterias basófilas estrictas. (10)

(ii) Mutantes de deleción debido a estructuras secundarias estables.

Es bien sabido que las plantillas de PCR que contienen estructuras secundarias estables a menudo producen una eficacia de amplificación o mutagénesis por deleción muy baja en los productos de PCR. Por ejemplo, los ARN ribosomales generalmente exhiben estructuras secundarias intensivas, la RT-PCR podría conducir a mutantes por deleción, que se excluirán del análisis posterior, ya que los genes de ARNr 16S amplificados a menudo se seleccionan por tamaño para así evitar la clonación o la electroforesis en gel de amplicones no específicos. (10)

Para evitar los problemas que surgen de las estructuras secundarias del molde durante el proceso de PCR de *Taq* polimerasa, Chou recomendó el uso de la proteína de unión al ADN monocatenario de *E. coli* en las reacciones de PCR o la aplicación de las ADN polimerasas, donde se describe que su capacidad de procesar es mayor que la de *Taq* Polimerasa. (10)

(iii) Mutantes puntuales debido a la incorporación errónea de ADN polimerasa.

Desde su primera aplicación en reacciones de PCR, se sabe que *Taq* Polimerasa presenta una tasa de incorporación errónea intrínseca durante la síntesis de cadenas, lo que puede conducir a sustituciones de bases. Según lo revisado por Eckert y Kunkel, las frecuencias de error observadas para la PCR basadas en Taq polimerasa pueden variar desde aproximadamente un error por cada 290 nucleótidos (3×10^{-3}) a un error por cada 5411 nucleótidos (2×10^{-4}), según en las condiciones de reacción utilizadas. Ford y col. Informaron de una tasa de error aún más baja de aproximadamente $2,6 \times 10^{-5}$ / pb por ciclo. Se midieron valores similares para la transcriptasa inversa. Stewart y col. Encontraron una tasa de incorporación errónea de un nucleótido por cada 700 bases para la ADN polimerasa, en la amplificación por PCR del genoma de ADN de 8 kilo bases del

virus del papiloma humano VPH 16. Varias ADN polimerasas termoestables contienen una actividad de exonucleasa (corrección de pruebas)

3 '→ 5', que da como resultado una tasa de incorporación errónea significativamente menor durante la síntesis de la hebra. (10)

Una tasa de error tan pequeña parece tener poco impacto en la evaluación filogenética de genes de ARNr 16S amplificados por PCR, ya que la tasa máxima de incorporación errónea daría lugar a cinco nucleótidos incorrectos para el gen completo (aproximadamente 1500 pb), lo que corresponde a una divergencia de secuencia del 0,3%. Sin embargo, si se realizan RT-PCR y / o varios ciclos de PCR, es decir, PCR anidada con cebadores específicos de grupo después de la amplificación específica de dominio, o los clones recombinantes se analizaron mediante amplificación por PCR del inserto y secuenciación del ciclo posterior con *Taq* polimerasa, la incorporación incorrecta puede acumularse dando lugar a una mayor tasa de error. La presencia de nucleótidos mal incorporados (o mal interpretados) es muy problemática cuando están ubicados en sitios que han sido seleccionados como diana de la sonda o cuando se utilizan pequeñas diferencias en la secuencia en la discriminación de cepas. La evaluación de la estabilidad de la estructura secundaria puede ayudar a identificar tales nucleótidos erróneos. (10)

- Variaciones de la secuencia del rRNA debido a la heterogeneidad del operón rrn.

Los genes de ARNr 16S de algunas bacterias y arqueas reflejan la aparición de rrn inter e intraespecíficas heterogeneidades del operón. Estas diferencias pueden interferir con el análisis de bibliotecas de clones de rDNA 16S o patrones de electroforesis en gel derivados de ecosistemas ambientales, ya que no está claro si una secuencia de rDNA 16S representa un organismo distinto o es solo un gen representativo de todo el operón rRNA 16S de un organismo. Debido a que es probable que los IVS se introduzcan en genes de ARNr 16S por transferencia lateral, su inclusión en análisis filogenéticos puede conducir a resultados erróneos. Por lo tanto, estas idiosincrasias de secuencia deben excluirse antes del análisis filogenético. Además, los IVS informados podrían conducir a la exclusión de los

respectivos rDNAs 16S ya que a la PCR a menudo le sigue la selección del tamaño de los amplicones para evitar análisis posteriores de productos no específicos. (10)

Pasos previos y posteriores a la PCR separados

1. Incluir controles negativos en los diferentes pasos de procesamiento de muestras y prueba de contaminación
2. Control de 'hisopo en blanco': hisopo estéril abierto en el laboratorio de secuenciación se somete a un protocolo de secuenciación completo
3. Control de 'extracción en blanco': extracción de ADN y todos los pasos posteriores se realizan sin material adicional
4. Control de 'biblioteca en blanco': no se aplica el protocolo de extracción y se utiliza agua libre de ADN como entrada de los pasos posteriores a la extracción para generar la biblioteca
5. Si la biomasa microbiana es baja, se pueden agregar controles adicionales en cuanto a la recolección de muestra
6. Si el resultado en términos de productos de PCR no es negativo, secuenciar el control negativo y restar esas secuencias computacionalmente.

Para minimizar el sesgo:

1. Reunir múltiples PCR para cada muestra
2. Preincubar con RNasas
3. Utilice polimerasas correctoras de errores, tiempos de hibridación más largos para reducir la formación de quimeras, potenciadores de PCR para mejorar los rendimientos.
4. Amplificación en un solo paso:
    a. Adecuado si se trabaja con un tipo de amplicón de muchas muestras
    b. No apto para amplificación de cebadores degenerados
    c. Protocolo rápido y poca pérdida de biomaterial
5. Amplificación en dos pasos:

a. Adecuado si se trabaja con muchos tipos de amplicones o metagenómica en un solo estudio.
b. Compatible con amplificación de cebadores aleatorios y degenerados
c. Protocolo más largo y más pérdida de biomaterial.

- La secuenciación puede generar sesgos que difieren entre métodos y laboratorios.

1. Utilice controles positivos: cepas puras de *Escherichia coli* para calibrar el rendimiento de la PCR y comunidades microbianas simuladas sintéticas que consisten en una mezcla de organismos de cultivo o ADN de organismos conocidos para verificar la posible introducción de sesgos o distorsiones durante el procesamiento de la muestra en la serie o la placa

2. Las secuencias de control pueden ser fácilmente discriminadas de las muestras experimentales mientras se procesan. Sin embargo, los controles son específicos para cada juego de cebadores así que se debieron rehacer para cada amplicón que se utilizó.

- Otras consideraciones para el procesamiento de muestras:

    - Aleatorización de muestras:

Las muestras deben ser aleatorias para controlar los efectos del lote y el efecto biológico de interés (por ejemplo, evitar procesar todas las muestras de control y todas las muestras de tratamiento en grupos separados).

    - Cobertura de secuencia

La cobertura necesaria depende del nivel de taxón en el que se realizará el análisis, la abundancia de la comunidad microbiana y cuán distintas son las muestras (una cobertura baja es suficiente para clasificar muestras muy diferentes, mientras que es necesaria una cobertura alta para discriminar muestras similares).

Análisis de los datos de la secuencia de ARNr 16S

El objetivo final de un análisis mediado por PCR de moléculas de ARNr 16S de microbiota compleja es la recuperación de información de secuencia, que permite la determinación de la diversidad microbiana, es decir, microorganismos cultivados y no cultivados, mediante análisis comparativo de secuencias de ARNr 16S. La secuenciación de genes de ARNr 16S amplificados y separados se puede realizar mediante técnicas estándar radiactivas o no radiactivas con cebadores de secuenciación universales y cebadores de rADN 16S internos.

La calidad de los resultados obtenidos mediante análisis comparativos de secuencias de ARNr 16S depende en gran medida del conjunto de datos disponible. Aunque se han liberado aproximadamente 5000 secuencias completas y parciales de ARNr 16S y rADN 16S de microorganismos cultivados y clones ambientales, este número refleja sólo una pequeña parte de la diversidad microbiana esperada. Los genes de ARNr 16S recuperados de muestras ambientales a menudo exhiben una baja similitud de secuencia con secuencias conocidas, lo que dificulta su afiliación filogenética. Esto lleva a la pregunta de si las secuencias ambientales representan microorganismos nuevos no cultivados o si no pueden asignarse a taxones conocidos debido al hecho de que, para muchos microorganismos cultivados, las secuencias de ARNr 16S no están disponibles o son de baja calidad (es decir, secuencias parciales y / o muchas secuencias ambiguas bases en secuencias liberadas).

En conclusión, se recomendó el uso combinado de todos los métodos disponibles para la detección de quimeras de ARNr, ya que un solo método, especialmente aquellos que se basan en el enfoque del vecino más cercano, no es suficiente. (10)

Comprobación cruzada de los resultados

El análisis de ARNr 16S mediado por PCR es una herramienta poderosa para la determinación de la diversidad microbiana de los ecosistemas ambientales. Aunque no existe una directriz general para un "buen análisis mediado por PCR del ARNr 16S de muestras ambientales", recomendaron la comparación de los resultados de

diferentes extracciones de ácidos nucleicos, amplificación por PCR y experimentos de clonación. Alternativamente, se podría usar una mezcla de ácidos nucleicos obtenida con diferentes métodos de extracción para múltiples amplificaciones por PCR y procedimientos de clonación, ya que es más probable que estos pasos introduzcan un error experimental. (10)

Análisis de Datos

Análisis del Metagenoma

- La preparación y el análisis de las muestras tiene un costo elevado y un procesamiento complejo y laborioso.
- La contaminación del ADN derivado del huésped puede ocultar las firmas microbianas.
- La profundidad de la secuenciación suele ser alta en comparación con otros métodos.
- No se discrimina entre vivos, muertos o activos.
- Los genomas microbianos promediados por la población tienden a ser inexactos debido a los artefactos de ensamblaje. (11)

Al momento de realizar el análisis de datos se cuenta con distintas versiones de software, estos pueden comportarse de diferente manera, por lo tanto:

1. Es necesario registrar la versión y los parámetros utilizados al momento de usar un software durante el análisis de datos. Las diferentes versiones del mismo software podrían tener diferentes parámetros predeterminados y errores no corregidos que podrían afectar el análisis.

2. Las versiones de paquetes utilizados por software específico, como R o Python, también deben documentarse.

Los resultados y las conclusiones del análisis deben ser coherentes, independientes del software y el método utilizados.

1. Aprovechar el *crowdsourcing* para la validación de resultados por investigadores externos utilizando sus propios métodos de análisis
   ○ Partiendo de los datos brutos o restringidos a pasos específicos del análisis de datos.
   ○ También se puede aplicar para evaluar la reproducibilidad de métodos experimentales.
   ○ Para verificar resultados y conclusiones biológicas de un estudio de metagenómica. (11)

Metagenómica

- Primero, el método de recolección debe preservar la firma microbiana de cada muestra.
- Segundo, las condiciones de recolección y almacenamiento de las muestras deben mantener la muestra biológica estable en condiciones de campo, durante el período de tiempo en el que se encuentre expuesta.
- En tercer lugar, las muestras recolectadas en un estudio prospectivo deben conservarse de una manera que maximice su potencial para su uso con múltiples plataformas (por ejemplo, microbiómica, metabolómica, transcriptómica).
- Por último, es probable que los estudios epidemiológicos futuros del microbioma necesiten ser muy amplios para ajustar las comparaciones múltiples y es posible que los datos deban agruparse o metaanálisis de múltiples estudios de diferentes laboratorios. transcriptómica) (9)

3.0 Planteamiento del Problema

Diferentes autores consideran que las mediciones de metataxonómica son muy propensas a sufrir sesgos, lo cual resulta en que las abundancias relativas, medidas de los taxa y genes en la muestra se distorsionan sistemáticamente de sus valores

reales. Se han estudiado diferentes fuentes de sesgo incluyendo diferentes cebadores de PCR que amplifican preferentemente diferentes conjuntos de taxa, diferentes protocolos de extracción pueden producir diferencias de 10 veces o más en la proporción medida de un taxón de la misma muestra y casi todas las elecciones en un experimento de metataxonómica han sido implicadas como contribuyentes al sesgo. Los ejemplos simulados han demostrado que el sesgo puede llevar a conclusiones cualitativamente incorrectas sobre qué taxa dominan diferentes muestras, qué ecosistemas son más similares, y qué taxa están asociados con una enfermedad determinada.

Una comprensión cuantitativa de cómo el sesgo distorsiona las mediciones de MGS también aclararía cómo los análisis y diagnósticos estadísticos se ven afectados por el sesgo y sugeriría alternativas más sólidas. (9)

Por lo cual se plantea la siguiente pregunta de investigación:

¿Los métodos de colección de las muestras, aislamiento de ADN bacteriano, procesamiento y análisis de datos influye sobre la calidad de las lecturas secuenciadas del gen 16S ARNr por secuenciación masiva paralela de la región hipervariable V3-V4 para la asignación taxonómica?

4.0 Justificación

Tras años de estudio es necesario detectar los errores en metataxonómica que probablemente conducirán a descripciones y resultados erróneos. El sesgo es una forma ubicua y particular de error sistemático que surge de las diferentes deficiencias con las que se miden varios taxa (es decir, preservados, extraídos, amplificados, secuenciados, o identificados y cuantificados bioinformáticamente) en cada paso, incluyen opciones relacionadas con la recolección de muestra, almacenamiento y preservación, extracción de ADN, cebadores amplificadores tecnología de secuenciación, longitud y profundidad de lectura, las diferentes bases de datos para alineación de secuencia y técnicas de análisis bioinformático, Cada paso físico, químico y biológico involucrado en el análisis molecular de un ambiente

es una posible fuente de sesgo que conducirá a resultados distorsionados, tomando en cuenta que alguno de los pasos aún no se encuentran estandarizados

5.0 Hipótesis

Al no cumplir con las buenas prácticas para los métodos de colección de las muestras, de aislamiento de ADN bacteriano, el procesamiento y análisis de datos entonces influirá negativamente en los estudios de metataxonómica provocando un sesgo al momento de la asignación taxonómica.

6.0 Objetivos

6.1 General

Analizar el comportamiento de calidad de las lecturas secuenciadas del gen 16S ARNr por secuenciación masiva paralela de la región hipervariable V3-V4 para la asignación taxonómica en función del método de colección y procesamiento de muestras y datos.

6.2 Específicos

1. Analizar cómo influye el tiempo de almacenamiento y preservación en la muestra de ADN.
2. Analizar la calidad de las lecturas en función del kit de extracción de ADN bacteriano.
3. Analizar la calidad de las lecturas con base en la integridad, pureza y concentración de ADN bacteriano.

7.0 METODOLOGÍA

7.1 Tipo de Estudio

Estudio analítico, observacional, transversal y retrospectivo.

7.2 Criterios de selección

Criterios de inclusión

1. Archivo FastaQ que cuente con secuencias forward y reverse.
2. Que cuente con datos de recolección (fecha de almacenamiento y condiciones de toma de la muestra)
3. Datos de concentración y pureza por fluorometría (Quibit)
4. Gel de agarosa al 1.5%
5. Kit de extracción de ADN bacteriano.
6. Reportes de calidad de las librerías

Criterios de exclusión

1. No hay criterios de exclusión

7.3 Estudio Preliminar

En el estudio se analizó un total de 40 muestras de las cuales 20 correspondieron a pacientes con adenocarcinoma pulmonar y 20 de controles sin cáncer, se extrajo el ADN genómico bacteriano, se realizó la secuenciación de amplicones de ADNr 16S (regiones hipervariables V3-V4) y se analizó la diversidad microbiana utilizando el programa CLC Genomics Workbench (QIAGEN CLC Microbial Genomics Module).

En los resultados del análisis de los datos de secuenciación, se observó una tendencia de diferenciación significativa entre los casos con adenocarcinoma pulmonar y el grupo control. Los géneros *Olsenella* (1.91 x10-3), *Pantoea* (p=0.03), y *Johnsonella* p=0.04 se enriquecieron en el grupo de cáncer de pulmón en comparación con el grupo de control y fueron estadísticamente significativos. Los datos sugieren que los pacientes con cáncer de pulmón son portadores de una comunidad de microorganismos diferente que los controles sanos.

En el análisis de las muestras se observó que una de las covariables que influyen en la composición bacteriana es la fecha de recolección de las muestras. El estudio

presenta limitaciones importantes principalmente en el programa de análisis (CLC *Genomics).*

7.4 Buenas Prácticas

Las buenas prácticas son la base para permitir la reproducibilidad y un mejor intercambio de los conjuntos de datos metagenómicos, lo que, en última instancia, conducirá a una mayor reutilización y publicación de los datos metagenómicos. Sólo mediante la adopción de normas y buenas prácticas se pueden evaluar en función de los principios de localización, accesibilidad, interoperabilidad y reutilización (*FAIR)* que deberían aplicarse a cualquier conjunto de datos científicos.

7.5 Toma de muestra salival, procesamiento y almacenamiento.

La recolección de muestras de saliva se llevó a cabo de acuerdo con el protocolo de Wong et al.

La toma de muestra se llevó a cabo por el personal involucrado en el proyecto, previamente entrenado, con el equipo de protección apropiado y con base en las buenas prácticas en investigación del microbioma humano de Knight y col. En un horario de entre las 9:00 y 11:00 hrs. Previo a la toma de muestra (una hora antes) el paciente debió evitar el consumir alimentos y/o bebidas, fumar, cepillar los dientes y besar, finalmente hacer la toma de muestra de aproximadamente 5 ml. En la cual el participante realizó con agua desionizada por 60 segundos, sentarse e inclinar su cabeza permitiendo la acumulación de la saliva en el piso de la boca y después escupir en el tubo Falcon de 50 ml. el cual fue rotulado con identificadores tal como establece Petra ten Hoopten. el cual siempre se mantuvo en un vaso con hielos aproximadamente a 4 °C en el cual no se exceda de 30 minutos para la colección de la saliva, muy importante y mencionado al paciente no toser, ni expectorar esputo.

La muestra fue colocada en un contenedor con refrigerantes para conservar la temperatura de 4°C, se centrifugó a 2600 g. durante un tiempo de 15 minutos para

permitir la separación del sobrenadante y el pellet. El pellet se separó en criovales y se añadieron 100 microlitros de PBS (buffer fosfato salino). Se asignó un número serial a cada uno de los viales para su identificación, el cual también fue colocado en la hoja de recolección de datos y todas las muestras fueron almacenadas a -80°C hasta su posterior uso.

7.6 Extracción, cuantificación e integridad del ADN bacteriano.

Para la extracción de ADN bacteriano se utilizó el kit *DNeasy UltraClean Microbial* Kit (QIAGEN), siguiendo el protocolo del proveedor con ciertas modificaciones de Rob Knight.

Una vez llevada a cabo la descongelación de las muestras a temperatura ambiente y el rotulado de los tubos a utilizar se tomaron 200 uL. de pellet y colocaron en un nuevo tubo Eppendorf de 2 mL. para centrifugar 100x100g por 10 minutos a temperatura ambiente, finalizando este tiempo se retiró el sobrenadante y colocar en los desechos, al nuevo pellet se le añadió 300 uL de solución *powerbeat* y 50 uL de solución SL, para incubar en *thermoblock* por 10 min a 55°C. posteriormente se colocó la muestra en el tubo con perlas, dando vórtex por 10 minutos, se volvió a centrifugar 100x100g por 2 minutos, se cambió el sobrenadante a un nuevo tubo de 2 mL, donde se le incorporaron 100 uL de solución IRS para nuevamente aplicar vórtex por 5 segundos; Se incubó a 4°C por 5 minutos, se volvió a centrifugar 100x100g por 2 minutos, se cambió el sobrenadante a un nuevo tubo de 2 mL, donde se incorporaron 100 uL de solución IRS para nuevamente aplicar vórtex por 5 segundos. Se incubó a 4°C por 5 minutos, se volvió a centrifugar 100x100g por 2 minutos y cambió nuevamente el sobrenadante en un nuevo tubo Eppendorf de 2 mL. Se agregaron 900 uL de solución SB y se dio vórtex por 5 segundos, se transfirió en un tubo con filtro, se centrifugó 100x200g por 2 minutos seguidos se colocaron 300 uL de solución CB y centrifugo 100x100g por 2 minutos, se eluyo con 40 uL de agua destilada libre de DNAsas (*ultraPure* de Invitrogen) precalentada a 55°C dejando reposar 5 minutos, se centrifugó 100x100g por 2 minutos para finalmente recuperar todo el volumen (aproximadamente 40 uL) colocando 6 uL en un tubo de

PCR de 0.2 mL. Y el restante en un criovial ya rotulado con su identificador para ser almacenado en -80°C.

Una vez obtenido el ADN bacteriano de cada muestra, se realizó la primera cuantificación con el equipo NANODROP 2000 (*Thermo Scientific*). Tomando un volumen de 2uL por muestra para su cuantificación (el cual tomó de los 6uL. Depositados en él tuvo de 0.2mL mencionados en la extracción del ADN bacteriano) y utilizando como blanco agua destilada libre de DNAsas/RNAsas (UltraPure de Invitrogen).

Para evaluar la integridad del material genético bacteriano se llevó a cabo la técnica de electroforesis en geles de agarosa al 1.5 % (invitrogen). Se utilizaron 100 mL de buffer Tris-acetato y EDTA (TAE) 1x (ThermoFisher). En todos los geles se designó el primer pozo para colocar el marcador compuesto por 1 uL de 1 kb DNA Ladder (Biolab), 1 uL de buffer de carga (Biolab) y 4 uL de agua destilada previamente filtrada. En los demás pozos se colocaron las muestras (4uL de cada muestra y 2mL de buffer de corrida.) Todos los geles se corrieron en la fuente de poder a 120V, durante 1 hora y se tiñeron con GelRed (PAGE GelRed).

7.7 Secuenciación

Para la preparación de librerías, se utilizó el protocolo 16S *Metagenomic Sequency Library preparation* (Ilumina). En primera instancia se realizó la amplificación de la hipervariable 3V-4V de la unidad bacteriana ribosomal pequeña 16s con los oligonucleótidos descritos en el protocolo.

16S Amplicon PCR Forward Primer = 5´

TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCCTACGGGGGCGCAG

16S Amplicón PCR Reverse Primer = 5´

GTCCGTGGGCTCGGAGATGTGTATAAGAGACAGGACTACGGGTA

Amplificación por PCR y secuenciación

Una vez hecha la mezcla de reacción por muestra la cual incluyó ADN microbiano, Amplicon PCR Forward Primer, Amplicon PCR Reverse Primer y 2x KAPA HiFi HotStart Ready Mix se colocaron los 25 uL en cada pozo de la placa o tiras de tubos (dependiendo del número de muestras). Se selló la placa o tiras de tubos (dependiendo del número de muestras). Se selló la placa de tubos y se llevó al termociclador para la amplificación.

Al concluir, se retiró del termociclador y se dio un spin para precipitar el líquido condensado; Como punto de control de calidad, 1 uL de productos de PCR se bioanalizó en el chip DNA 1000 para verificar el tamaño de fragmentos (usando los cebadores V3 y V4 del gen 16S bacteriano, se obtuvieron amplicones de 550 pb).

7.8 Análisis Bioinformático

El análisis de las muestras se llevó a cabo en el software RStudio con la paquetería DADA2. El primer pasó fue ingresar las secuencias en formato *fastaq paired-end* de las 40 muestras seleccionada para posteriormente llevar a cabo la clasificaron en secuencias Forward y Reverse para realizar la inspección de calidad de las mismas con las librerías ggplot2 y *cowplot*, una vez obtenidos los gráficos de calidad de las muestras y teniendo claro los puntos de corte, se realizó el corte 280 Forward 220 Reverse, se verificaron los cortes por medio tasas de error y así evitar algún tipo de sesgo en pasos posteriores, lo siguiente fue el realizar el emparejamiento de las secuencias (Forward y Reverse) y así obtener un resumen de cada una de las muestras, se detectaron secuencias quiméricas para su eliminación posterior y obtener el porcentaje de lecturas verdaderas. Una vez terminado el pre-procesamiento de las 46 se llevó a cabo la asignación taxonómica por medio de la base de datos 16S SILVA V.138.1. Una vez hecha la asignación se ingresó la metadata con las variables a trabajar y así se obtuvo la significancia a nivel filogenético (filo y género).

7.9 Variables

TABLA 1 DESGLOSÉ DE TODAS LAS VARIABLES

Variable	Tipo de variable	Definición conceptual	Definición operacional	Escala de Medición	Medidas de tendencia central	Medidas de dispersión	Forma	Representación gráfica
Recolección de muestras	Nominal	Según el protocolo de buenas prácticas el método de recolección debe preservar la firma microbiana de cada muestra (horario específico, saliva basal, no más de 30 minutos, estandarización o dominio de la técnica , procedimientos consistentes por parte del paciente, conservación de la muestra durante el proceso).	Que cumpla los criterios establecidos en las buenas prácticas.	Cumple No cumple	Frecuencia absoluta (fa) Frecuencia relativa (fr) Porcentaje (%)	No aplica	No aplica	Diagrama de barras Scores circulares (pie) ideal para 3-5 categorías
Almacenamiento y preservación de muestras.	Nominal	Conforme a las buenas prácticas el almacenamiento a -80°C es un protocolo estándar, pero el impacto del almacenamiento de muestras a largo plazo, así como el proceso de congelación y descongelación de muestras en la integridad del microbioma sigue siendo una cuestión abierta.	Que cumpla con los criterios establecidos en las buenas prácticas. (temperatura, sustancias estabilizadoras, ciclos de congelación-descongelación y tiempo de almacenamiento).	Cumple No cumple	Frecuencia absoluta (fa) Frecuencia relativa (fr) Porcentaje (%)	No aplica	No aplica	Diagrama de barras Scores circulares (pie) ideal para 3-5 categorías
Lote de Kit de extracción	Nominal	Protocolos específicos de muestra y kits comerciales que maximicen el rendimiento, la calidad y la integridad de la extracción de ADN de la estructura y diversidad de la comunidad microbiana. Si la biomasa microbiana es baja, se pueden agregar controles adicionales en cuanto a la recolección de muestra.	De acuerdo con el protocolo, las muestras deben seguir el protocolo del fabricante, hacer una buena elección del kit y utilizar el mismo procedimiento durante todo el estudio.	Extracción con mismo lote de kit. Extracción con diferente lote de kit.	Frecuencia absoluta (fa) Frecuencia relativa (fr) Porcentaje (%)	No aplica	No aplica	Diagrama de barras Scores circulares (pie) ideal para 3-5 categorías
Calidad de DNA.	Nominal	La cantidad y calidad de ADN de un organismo que está disponible para formar parte de una biblioteca de secuenciación depende de la eficiencia del protocolo de extracción de ADN Las amplificaciones por PCR pueden generar sesgos cuando la concentración de la plantilla de ADN es baja o el número de ciclos de PCR es alto.	Que cumpla con los criterios de cantidad, calidad e Integridad de ADN dentro de los parámetros establecidos.	La cantidad, calidad e integridad de ADN es la adecuada. La cantidad, integridad y calidad de ADN no es adecuada.	Frecuencia absoluta (fa) Frecuencia relativa (fr) Porcentaje (%)	No aplica	No aplica	Diagrama de barras Scores circulares (pie) ideal para 3-5 categorías.
Concentración de ADN bacteriano	Numérica continua	La concentración de ADN está basada en la absorbancia de un compuesto presente en una longitud de onda determinada.	Concentración de ADN superior de 20 ng/µl	Se medirá en Ng/ml	Media Moda Mediana	Desviación estándar Rango o amplitud		Diagramas diferenciales,diagramas integrales y gráficos
Concentración de trabajo	Numérica continua	Concentración de ADN bacteriano mínima para ser incluida en el estudio.	Concentración de ADN superior de 20 ng/µl	Cumple con concentración mínima No cumple con concentración mínima	Media Moda Mediana	Desviación estándar Rango o amplitud		Diagramas diferenciales,diagramas integrales y gráficos
Integridad de ADN bacteriano	Nominal	Se comprueba la integridad de la muestra de ADN mediante	Mediante el proceso de electroforesis si la muestra	ADN no degradado.	Frecuencia absoluta (fa)	No aplica	No aplica	Diagrama de barras

		electroforesis en gel. Si una muestra de ADN aparece degradada, la muestra se retira y se destruye.	se encuentra fuera de parámetros y degradada no se incluirá en el estudio.	ADN degradado.	Frecuencia relativa (fr) Porcentaje (%)			Scores circulares (pie) ideal para 3-5 categorías
Integridad de ADN bacteriano de trabajo	Nominal	La muestra de ADN bacteriano no deberá estar degradada y mantener su integridad para ser incluida en el estudio.	Mediante el proceso de electroforesis si la muestra se encuentra fuera de parámetros y degradada no se incluirá en el estudio.	ADN no degradado. ADN degradado	Frecuencia absoluta (fa) Frecuencia relativa (fr) Porcentaje (%)	No aplica	No aplica	Diagrama de barras Scores circulares (pie) ideal para 3-5 categorías
Pureza	Numérica continua	La pureza de las muestras se evalúa mediante la relación de absorbancia, utilizando un espectrofotómetro respetando parámetros estrictos.	Relación de pureza del ADN de 1,7-2,0.	Cumple con la relación de pureza No cumple con la relación de pureza.	Media Moda Mediana	Desviación estándar Rango o amplitud	No aplica	Diagramas diferenciales,diagramas integrales y gráficos
Pureza de trabajo	Numérica continua	Pureza mínima de ADN para ser incluida en el estudio.	Relación de pureza del ADN de 1,7-2,0.	Cumple con la relación de pureza No cumple con la relación de pureza.	Media Moda Mediana	Desviación estándar Rango o amplitud		Diagramas diferenciales,diagramas integrales y gráficos
Calidad de ADN de trabajo	Nominal	La muestra deberá cumplir con la pureza, integridad y cantidad mínima de ADN para ser incluida en el estudio.	La muestra de ADN deberá cumplir con una relación de pureza de 1,7-2,0, concentración superior de 20 ng/µl y no estar degradada.	La cantidad, calidad e integridad de ADN es la adecuada. La cantidad, integridad y calidad de ADN no es adecuada.	Frecuencia absoluta (fa) Frecuencia relativa (fr) Porcentaje (%)	No aplica	No aplica	Diagrama de barras Scores circulares (pie) ideal para 3-5 categorías
Calidad de lecturas en los análisis secundarios.	Nominal	Es importante visualizar o tener conocimiento de las secuencias aptas para poder llevar a cabo una asignación taxonómica de calidad y así evitar sesgos por posibles al realizar la asignación.	Se seleccionarán las secuencias de calidad para evitar sesgos y realizar una asignación taxonómica adecuada.	Cumplió con filtro de calidad de lecturas. No cumplió con el filtro de calidad de lecturas.	Frecuencia absoluta (fa) Frecuencia relativa (fr) Porcentaje (%)	No aplica	No aplica	Diagrama de barras Scores circulares (pie) ideal para 3-5 categorías
Base de datos de referencia	Nominal	Mediante un análisis comparativo de secuencias de ARNr 16S se realiza la recuperación de información de secuencia, que permite la determinación de la diversidad microbiana. Actualmente no hay estándares.	A la actualidad existen diferentes bases de datos de 16S las cuales llevan a cabo actualizaciones de las mismas a diferentes tiempos por lo cual es necesario la revisión de las mismas	-GREEN GENES -SILVA -RDP	Frecuencia absoluta (fa) Frecuencia relativa (fr) Porcentaje (%)	No aplica	No aplica	Diagrama de barras Scores circulares (pie) ideal para 3-5 categorías

8.0 Resultados

8.1 Análisis Descriptivos

En el presente estudio se realizó el análisis de 50 muestras pertenecientes a pacientes con adenocarcinoma pulmonar y pacientes control sin adenocarcinoma pulmonar para poder determinar las posibles variables que pueden influir sesgando los estudios del gen 16S ARNr en función del tiempo de almacenamiento y preservación de la muestra, analizando la calidad de las lecturas en función del lote de kit de extracción y analizando la calidad de las lecturas con base en la integridad, pureza y concentración de ADN bacteriano.

El primer punto que se analizó, fueron las características en las cuales se recolectaron las muestras (horario, tiempo de recolección, y almacenamiento), no se encontró alguna diferencia significativa entre ellas, siendo que las 50 muestras contaban con buenas prácticas con respecto a los métodos de recolección.

Posteriormente se revisaron las técnicas de aislamiento del material genético bacteriano, se encontró que 19 muestras contaban con extracción del lote de kit Qiagen 157042821, las cuales correspondían al 38% y 31 muestras contaban con lote de kit Mobio BS16C9. Las cuales correspondían al 62%. Se siguieron los protocolos específicos de cada kit para maximizar la calidad e integridad de la extracción, utilizando el mismo procedimiento en cada una de las muestras.

Para los pasos posteriores se clasificaron en calidad alta y baja para así contar con mayor control de las mismas. 49 muestras correspondientes al 98% contaron con una calidad alta mientras el 2% correspondiente a una muestra contaba con una calidad baja.

Se llevó a cabo el análisis de datos tomando en cuenta la concentración, integridad y pureza de ADN; Donde la concentración debió ser superior de 20 ng/µl, se evaluó si cumplían o no con dicho criterio, obteniendo como resultado que 49 de las muestras equivalente al 98% cumplieron, mientras que 1 muestra correspondiente al 2% no cumplió con dicho parámetro. La integridad del ADN se comprobó

mediante geles de agarosa donde se pudo encontrar si alguna de las muestras contaba con degradación, encontrando que 19 de las muestras correspondientes al 38% se clasificaron como material genético bacteriano íntegro y 31 muestras con el 62% se clasificaron como intermedio. Por último, se evaluó la pureza de trabajo tomando en cuenta la relación 260/280 y 260/230 como parámetro óptimo en el primero de 1.7 - 2.1 y el segundo de 1.8 - 2.2, evaluando como media, alta y baja, donde se obtuvo que 50% de las muestras se clasificaron como alta, 30% media y 20% baja.

Finalmente se evaluó la calidad del ADN bacteriano. Si las muestras cumplían con los tres criterios antes mencionados (concentración, integridad y pureza) se consideró calidad alta, si contaban con solo dos se clasificó como calidad media y si no contaron con alguno de los tres se les denominó calidad baja, de la cual no obtuvimos muestra alguna. Aplicando la prueba estadística correspondiente se encontró diferencia estadísticamente significativa con una p=0.00007501.

8.2 Análisis Estadístico

Tabla 2 Análisis Estadístico

Variable				**p**
Fecha de recolección	**2008-2014** 39 (78%)	**2015-2018** 11 (22%)		0.57
Kit de extracción	**Qiagen** **157042821** 19(38%)	**Mobio** **BS16C9** 31(62%)		0.08
Concentración ADN (Nanodrop)	69.8770.73 (6.4-250.4)			-
Concentración ADN (Qubit)	38.76+-31.17 (2.155-124.950)			-
Calidad de ADN	**Alto** 39 (78%)	**Medio** 11 (22%)		0.00007501
Concentración de trabajo	**Cumple** 49 (98%)	**No cumple** 1 (2%)		-
Integridad	**Integro** 19 (38%)	**Intermedio** 31 (62%)		0.08969
Pureza de trabajo	**Alto** 25 (50%)	**Medio** 15 (30%)	**Bajo** 10(20%)	0.0302
Librerías Nanomoles	117.2 +- 49.8403 (2.9-276.2)			-
Calidad de las librerías	**Alta** 49 (98%)	**Baja** 1 (2%)		-

8.3 Análisis primario

En esta sección se verificaron las métricas del *Phred Quality Score*, densidad de clusters, porcentaje de *clusters* filtrados, número de lecturas y el porcentaje de alineamiento.

1. *Phred Quality Score*: Con base en las guías de Ilumina y el kit de secuenciación utilizado, el 70% de los datos secuenciados debe tener un score mayor de 30, los datos obtenidos mostraron que el 64.4% de los datos tuvieron un score mayor de 30.

2. Densidad de clúster: Es la densidad de grupos (en miles por mm^2) detectada por análisis de imágenes, +/- 1 desviación estándar, el resultado obtenido fue de 658 ± 21.

3. Porcentaje de *clusters* filtrados: De un total de 33,534,276 lecturas, el 92.76% de ellas lograron pasar el filtro de calidad (31,095,876 lecturas).

4. Número de lecturas: 33,534,276 fue nuestro número de lecturas totales.

Porcentaje de alineamiento: Corresponde al porcentaje de la muestra que se alineó con el genoma *PhiX*, que se determina para cada nivel o lectura independiente, se obtuvo el 22.93%. Finalmente, el sistema de Illumina nos proporciona el estatus general de la calidad de la corrida realizada con base en los criterios antes mencionados, el estatus general de la corrida de secuenciación fue “bueno”, por lo tanto, se prosiguió con el análisis secundario.

READ	CYCLES	YIELD	PROJECTED YIELD	ALIGNED (%)	ERROR RATE (%)	INTENSITY CYCLE 1	%>Q30
Read 1	301	4.66 Gbp	4.66 Gbp	23.65	3.17	67.71	67.76
Read 2 (I)	8	108.84 Mbp	108.84 Mbp	0.00	0.00	281.03	68.60
Read 3 (I)	8	108.84 Mbp	108.84 Mbp	0.00	0.00	202.08	87.61
Read 4	301	4.66 Gbp	4.66 Gbp	22.21	4.33	54.89	59.67
Non-index Reads Total	602	9.33 Gbp	9.33 Gbp	22.93	3.75	61.30	63.72
Total	**618**	**9.55 Gbp**	**9.55 Gbp**	**22.93**	**3.75**	**151.43**	**64.04**

ILUSTRACIÓN 1 ANÁLISIS PRIMARIO DE SECUENCIACIÓN

8.4 Análisis Secundario

Pre-procesamiento.

Se ingresaron las secuencias de 50 muestras para su análisis, los primeros pasos fueron el realizar la separación de las secuencias conforme a *forward* y *reverse,* esto con la finalidad de ser precisos en la calidad de las mismas.

Calidad de ADN

Una vez ingresadas las secuencias se realizaron los gráficos de calidad (figura 2), estableciendo como puntos de corte 280 *forward* (figura 2A) y 190 *reverse* (figura 2B), a lo mismo recortando de los extremos 5´ 19 nucleótidos correspondientes a los adaptadores.

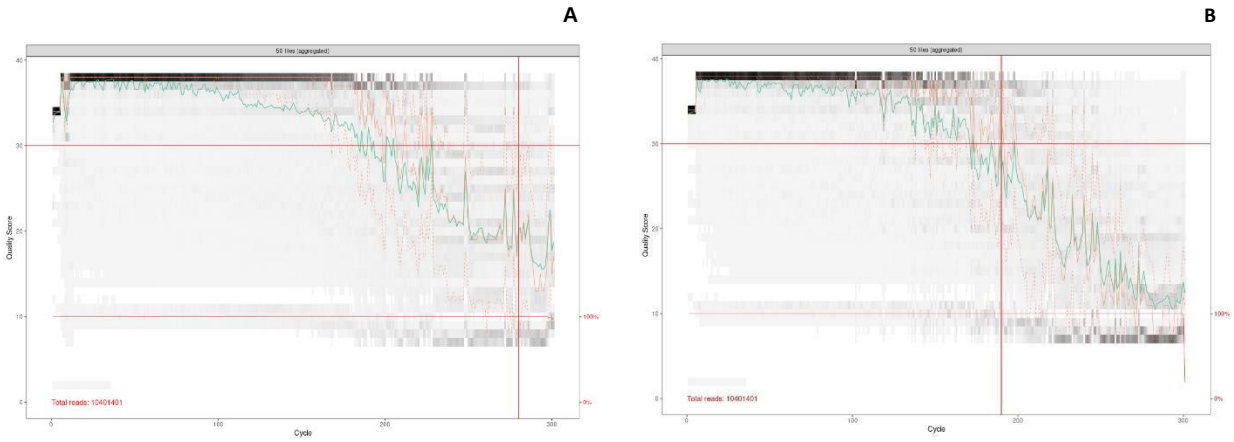

ILUSTRACIÓN 2 CALIDAD DE MATERIAL GENÉTICO BACTERIANO DE REGIONES HIPERVARIABLES V3-V4, CALIDAD FORWARD (A) REVERSE (B)

Tasas de Error

Posteriormente se obtuvieron las tasas de error, esto con la finalidad de solucionar los probables errores al momento de la secuenciación y así poder obtener resultados con mayor precisión al momento de la asignación taxonómica.

En la figura 3A se observa las tasas de error que presentaron las lecturas *forward* y en la figura 3B para lecturas *reverse*, en ambas figuras la línea roja nos indica la manera adecuada en que las lecturas deberían comportarse, esto con la finalidad de mostrar que entre mayor sea la calidad menor fueron los errores ocurridos. Los círculos negros corresponden a las lecturas de las muestras, los cuales siguieron muy de cerca el patrón sugerido.

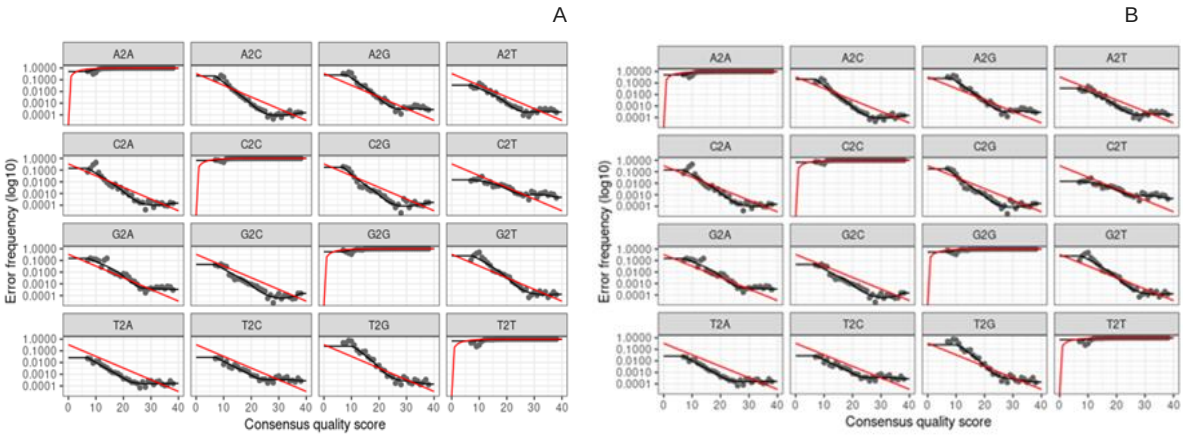

ILUSTRACIÓN 3 TASAS DE ERROR FORWARD (A) Y REVERSE (B)

8.5 Asignación taxonómica

Una vez obtenidos estos resultados se realizó la asignación taxonómica de 4742 secuencias obtenidas de las 50 muestras procesadas, por medio de la base de datos SILVA versión 138.1, posterior a ello se eliminaron aquellas secuencias correspondientes de eucariotas, mitocondria y cloroplastos, para así evitar sesgos al momento del análisis diferencial. Una vez realizada la asignación y filtrado se obtuvo un total de 723 taxa en las 50 muestras, esto en 7 niveles taxonómicos.

Diversidad Alfa de acuerdo al Año de Recolección

Se analizó el año de recolección de las muestras dividiéndolos en dos intervalos el primero incluyó las muestras recolectadas de 2008 a 2014 y el segundo intervalo fue de 2015 a 2018, obteniendo 39 muestras en el primer grupo y 11 correspondientes al segundo intervalo. Posteriormente se obtuvo la diversidad alfa para ver si había alguna diferencia entre los grupos de dicha variable, (figura 9). Se utilizó el índice de Shannon en el cual se obtuvo un valor de p=0.72 mientras que en el índice de Simpson se obtuvo un valor de p=0.36, demostrando que no había diferencias significativas entre estos dos grupos.

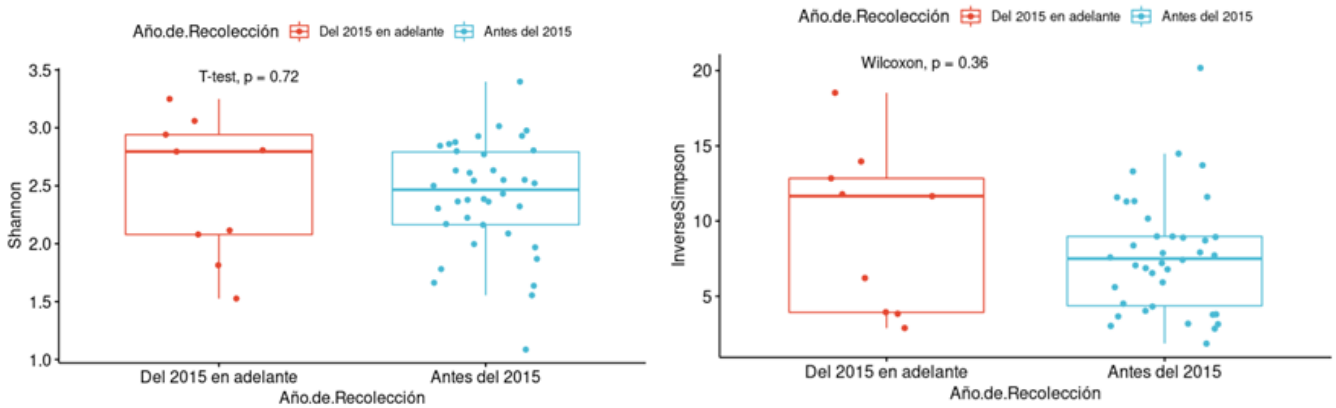

ILUSTRACIÓN 4 DIVERSIDAD ALFA RESPECTO AL AÑO DE RECOLECCIÓN, DIVIDIDO EN INTERVALOS DEL 2008 - 2014 Y 2014-2018, CON UNA P= 0.72

Diversidad Beta conforme al Año de Recolección

Se realizó el análisis de coordenadas principales (PCoA), basado en la distancia UniFrac no ponderada y la distancia UniFrac ponderada con la finalidad de observar la existencia de relación de la estructura microbiana de los grupos de manera cuantitativa y cualitativa con la cual no se observó diferencia o alguna aglomeración específica de la misma.

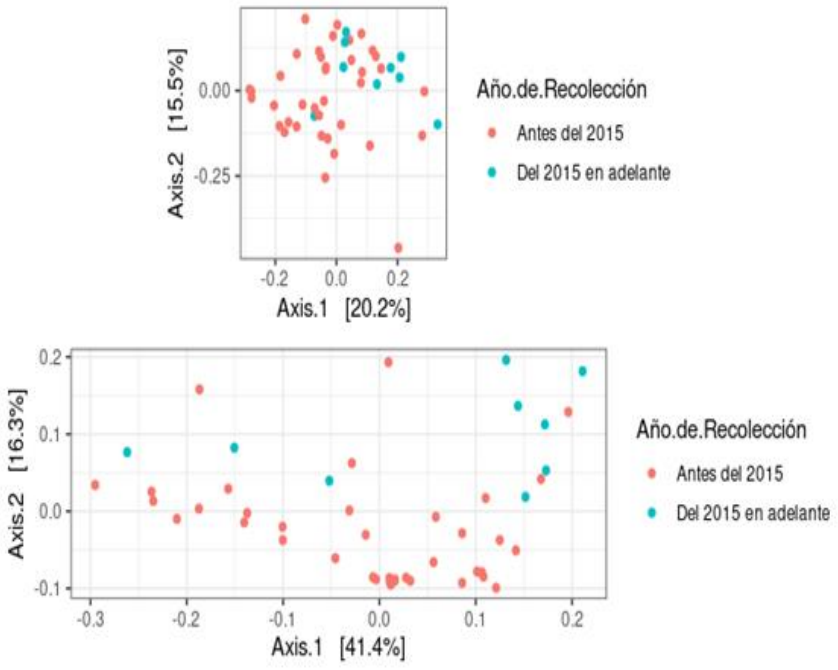

ILUSTRACIÓN 5 DIVERSIDAD BETA DEL AÑO DE RECOLECCIÓN, DONDE SE DESCRIBE CUALITATIVA Y CUANTITATIVAMENTE SIN PRESENTAR AGLOMERACIONES.

Abundancia Relativa conforme al Año de Recolección

Se obtuvo la abundancia en las muestras recolectadas antes del 2015 y después del 2015 de acuerdo al filo (A) y al género (B). En la figura A se obtuvo la abundancia, donde predominó el filo *Actinobacteriota* con 62.5% antes del año 2015 y 48.4% después del año 2015, en el cual se hicieron presentes 10 variantes de la misma, seguido por *Firmicutes* con 18% y 23.1% de 2015 en adelante, *Patescibacteria* con 14.1% antes de 2015 y 7.7% después del 2015. En la figura B, el género *Rothia* presentó una mayor abundancia obteniendo un porcentaje de 39.8% antes del año 2015 y 27.2% después del 2015, el cual presentó 6 diferentes variantes de la misma, seguido por *Actinomyces* con 9.5% y después de 2015 13.3%, TM7x con 8.7% y 5.6% después de 2015.

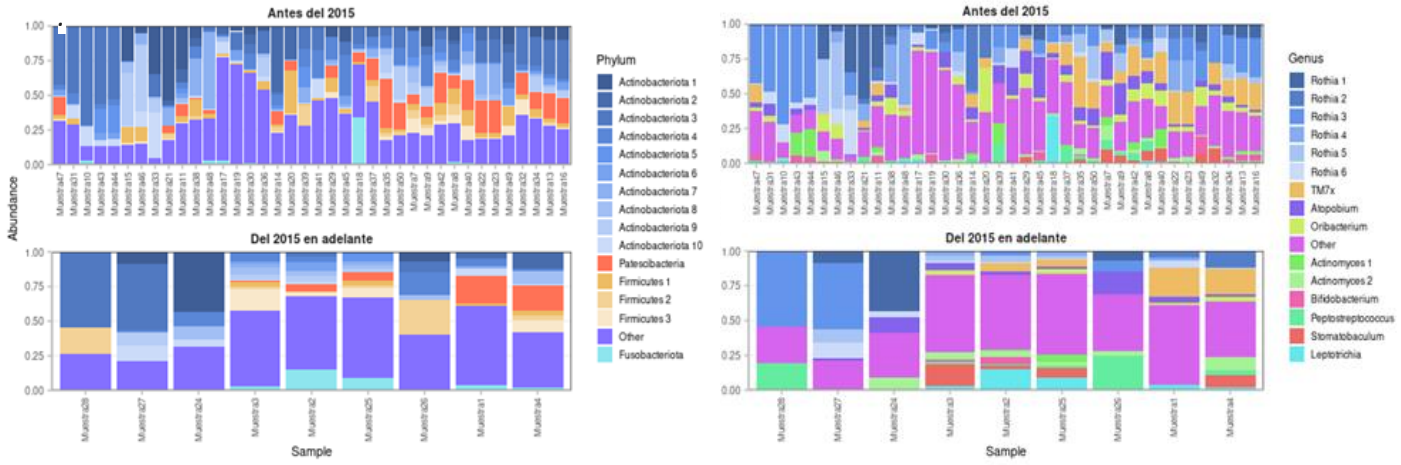

ILUSTRACIÓN 6 ABUNDANCIA RELATIVA CONFORME AL AÑO DE RECOLECCIÓN PREDOMINANDO ROTHIA A NIVEL DE GÉNERO Y ACTINOBACTERIOTA A NIVEL DE FILO.

Análisis Diferencial respecto al Año de Recolección

Se obtuvo la diferencia entre el año de recolección donde se encontraron aumentados los géneros de *Mogibacterium, Brucella. Oribacterium, Actinomyces y Candidatus Saccharimonas* y a nivel de filo se encontraron aumentados *Patescibacteria, Actinobacteriota, Firmicutes y Proteobacteria.* Encontrando una diferencia estadísticamente significativa.

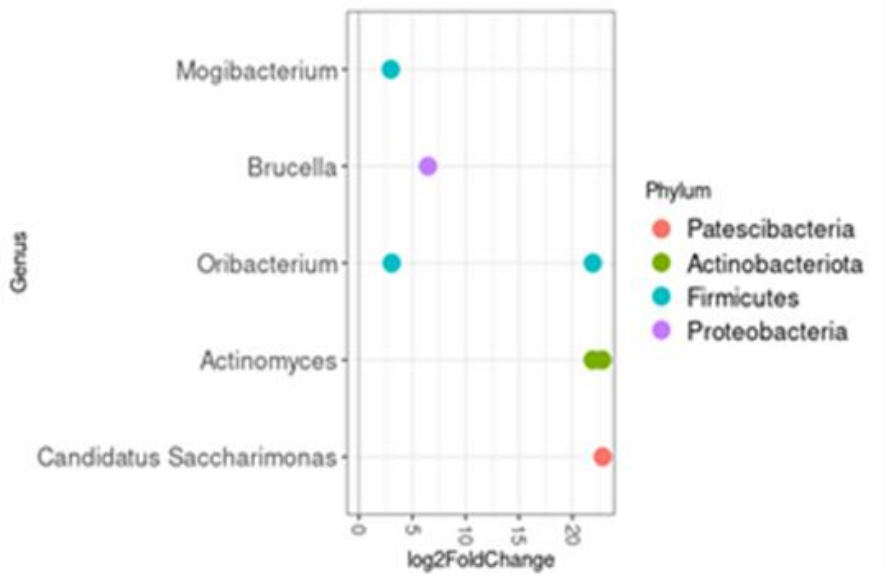

ILUSTRACIÓN 7 EL ANÁLISIS DIFERENCIAL RESPECTO AL AÑO DE RECOLECCIÓN MUESTRA UNA DIFERENCIA SIGNIFICATIVA

Diversidad Alfa respecto al Kit de Extracción

Se analizó el kit de extracción que se utilizó para el procesamiento de las 50 muestras usando dos diferentes kits, Qiagen 157042821 y Mobio BS16C9. En la figura A, se realizó el índice de Simpson obteniendo un valor de p=0.49 donde 19 muestras pertenecen al kit Qiagen 157042821 y 31 correspondientes al kit Mobio BS16C9 Qiagen 157042821 y Mobio BS16C9. En el índice de Shannon (figuró B), se obtuvo un valor de p igual a 0.86, 17 muestras pertenecieron a Qiagen 157042821 y 30 muestras a Mobio BS16C9.

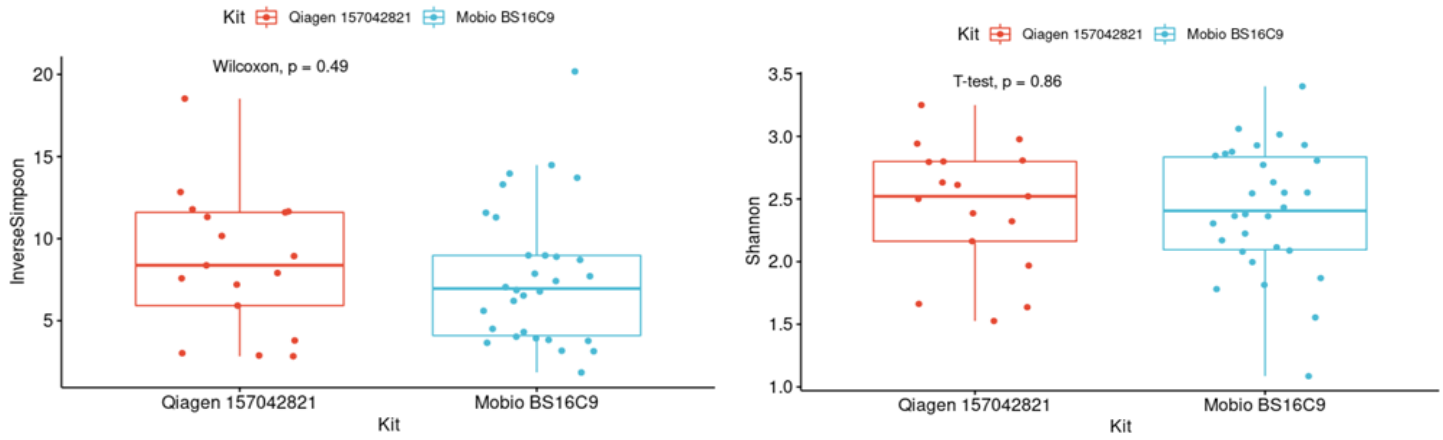

ILUSTRACIÓN 8 DIVERSIDAD ALFA RESPECTO AL KIT DE EXTRACCIÓN, QIAGEN 157042821 Y MOBIO BS16C9 CON UN VALOR DE P=0.49

Abundancia Relativa respecto al Kit de Extracción

Los siguientes gráficos mostraron la abundancia respecto al Kit de extracción que se utilizó para las 50 muestras, clasificándolos en Mobio BS16C9 y Qiagen 157042821, de acuerdo al Género (A) y Filo (B) donde obtuvimos que a nivel de género *Rothia* fue la más abundante con 59.9% en el kit Mobio BS16C9 y 59.5% en Qiagen 157042821, mostrando 6 diferentes variantes de la misma, seguida por *Actinomyces*, en ambos kits de extracción, mientras que a nivel de Filo *Actinobacteriota* fue la más abundante con 59.9% en el kit Mobio BS16C9 y 59.5% en Qiagen 15704282, seguida por *Firmicutes*, presentando 10 diferentes variantes, los cuales se pueden apreciar en la segunda gráfica.

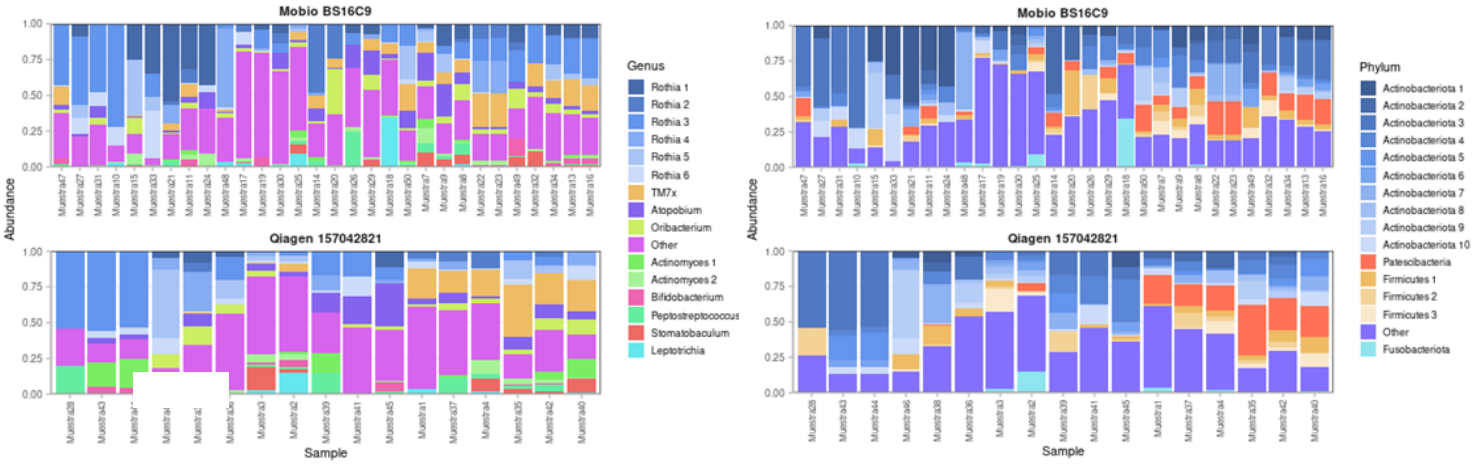

ILUSTRACIÓN 9 ABUNDANCIA RELATIVA RESPECTO AL KIT DE EXTRACCIÓN, PREDOMINANDO ROTHIA A NIVEL DE GÉNERO Y ACTINOBACTERIOTA A NIVEL DE FILO.

Se realizó el diagrama de Venn para tener una mejor visión de las taxas en donde no se obtuvo ninguna diferencia estadísticamente significativa, obteniendo 0 % en todas las áreas

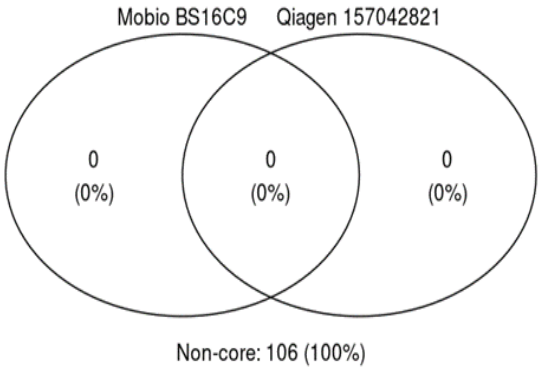

ILUSTRACIÓN 10 DIAGRAMA DE VENN PARA KIT DE EXTRACCIÓN MOBIO BS16C9 Y QIAGEN 157042821, DONDE NO SE OBTUVO NINGUNA DIFERENCIA

Análisis Diferencial respecto al Kit de Extracción

Se obtuvo el análisis de diferenciación en los kits de extracción en el cual se encontró a nivel de Filo a *Proteobacteri*a y *Firmicutes* disminuidos y a nivel de género a *Orinobacterium* y *Brucella,* igualmente disminuidos, no encontrando una diferencia estadísticamente significativa.

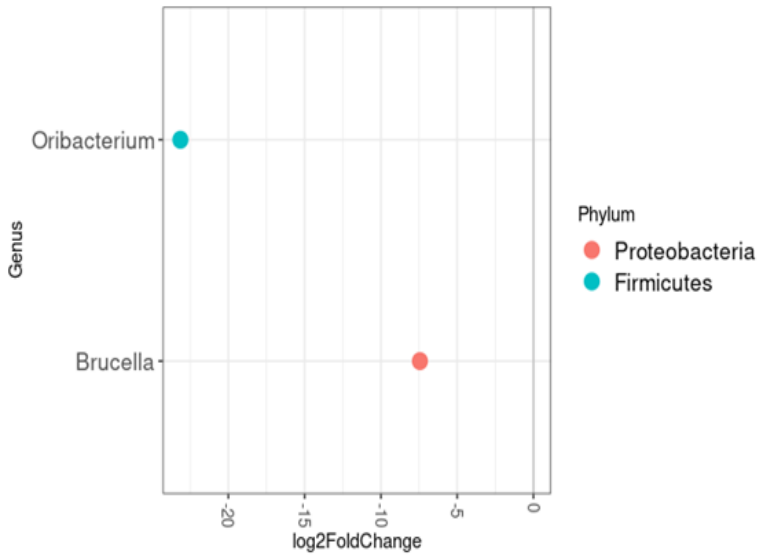

ILUSTRACIÓN 11 EL ANÁLISIS DIFERENCIAL RESPECTO AL KIT DE EXTRACCIÓN, MUESTRA FILO Y GÉNERO DISMINUIDOS, SIN PRESENTAR UNA DIFERENCIA ESTADÍSTICAMENTE SIGNIFICATIVA.

Abundancia de acuerdo a la Calidad de Librerías

Los siguientes gráficos muestran la abundancia de acuerdo al Filo (A) y al Género (B) respecto a la calidad de las librerías, clasificándolas en dos categorías, calidad de librerías alta y calidad de librerías baja, donde únicamente la muestra 11 presentó una calidad baja.

En la figura A, se obtuvo una abundancia predominantemente de *Actinobacteriota* en calidad alta, la cual presentó diez variantes de esta, mientras que en calidad baja para la muestra 11 los resultados fueron similares predominando *Actinobacteriota*.

En la figura B, se obtuvo la abundancia predominando *Rothia* presentando y 6 variantes de esta en calidad alta, mientras que en la muestra 11 correspondiente a calidad baja se obtuvo que *Rothia* predominó de igual manera, seguido de *TM7x*.

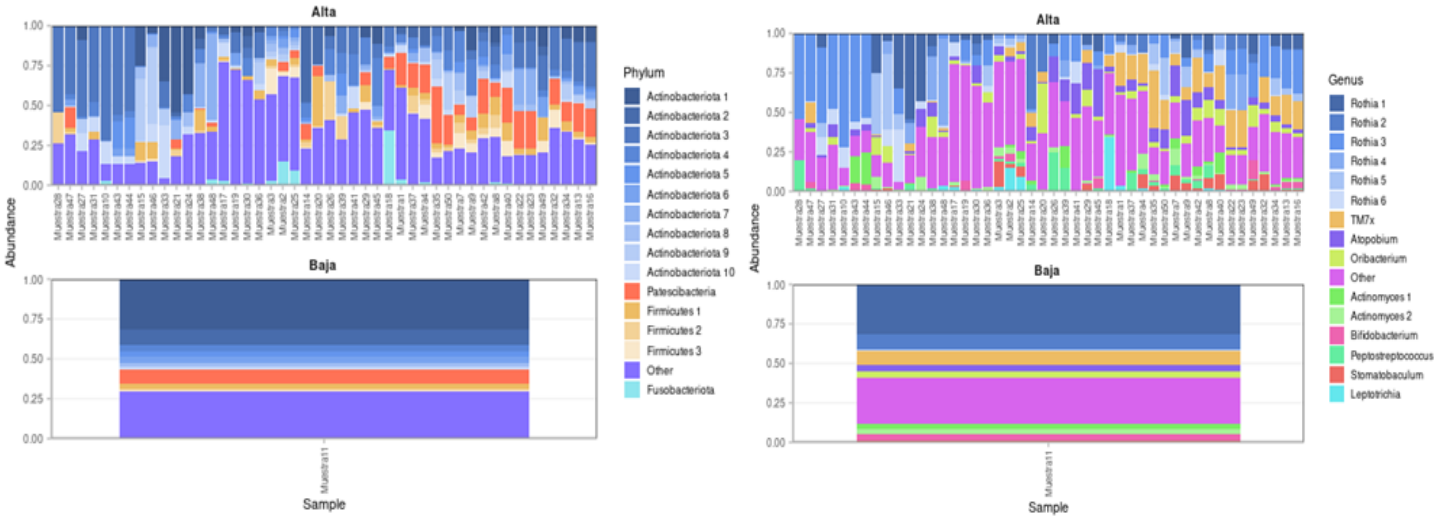

ILUSTRACIÓN 12 ABUNDANCIA RELATIVA RESPECTO A LA CALIDAD DE LIBRERÍAS.

Integridad de Trabajo

Abundancia de acuerdo a la Integridad de Trabajo

Se obtuvo la abundancia respecto a la Integridad de trabajo clasificándose en ADN no degradado y ADN degradado. En la figura A se obtuvo la abundancia respecto al filo y en la figura B respecto al género. Teniendo como resultado que a nivel de filo *Actinobacteriota* fue la más abundante con un 59.8% y se encontraron diez subtipos de la misma, mientras que a nivel de Género predominó *Rothia* con un 37.4% 6 subtipos de esta, seguida por *Actinomyces*.

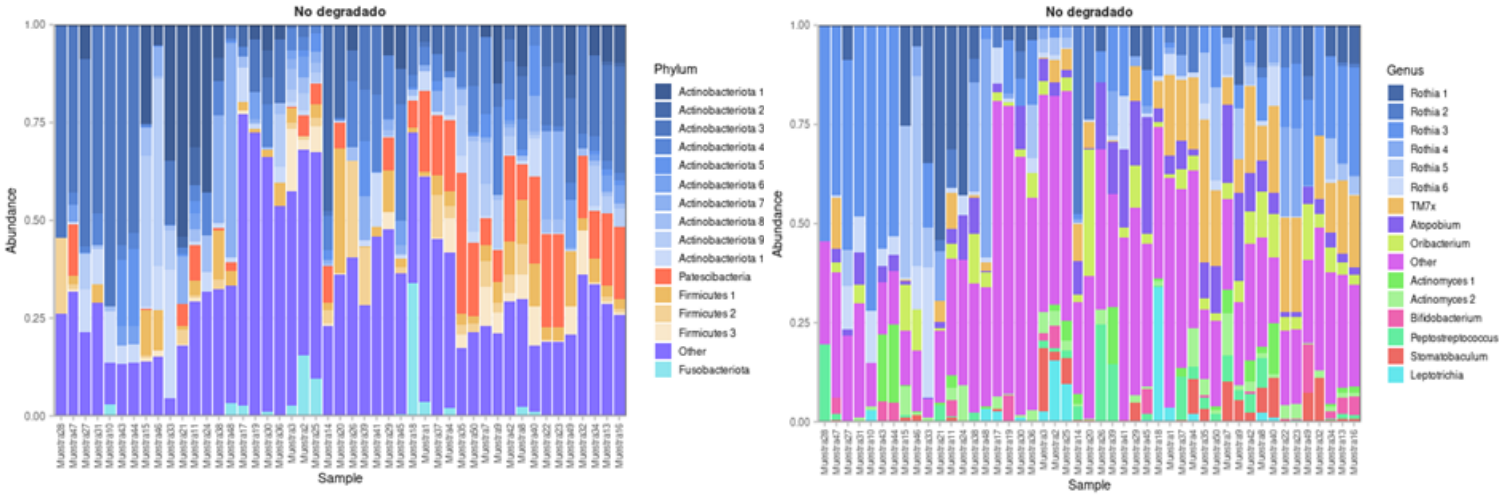

ILUSTRACIÓN 13 ABUNDANCIA RELATIVA RESPECTO A LA INTEGRIDAD DE TRABAJO, PREDOMINANDO ROTHIA A NIVEL DE GÉNERO Y ACTINOBACTERIOTA A NIVEL DE FILO.

Concentración de trabajo

Abundancia Relativa de acuerdo a la Concentración de Trabajo

Se obtuvo la abundancia relativa respecto a la concentración de trabajo donde de las 50 muestras que se analizaron, la única que no cumplió con los parámetros de concentración fue la muestra 19. La imagen A muestra la abundancia relativa respecto al género y la imagen B respecto al filo.

Teniendo como resultado que a nivel de género *Rothia* fue la más abundante presentando seis variantes de la misma, seguida por *TM7x* y *Atopobium*, mientras que a nivel de Filo *Actinobacteriota* fue la más abundante presentando 10 variantes de la misma, seguida por *Patescibacteria* y *Firmicutes*.

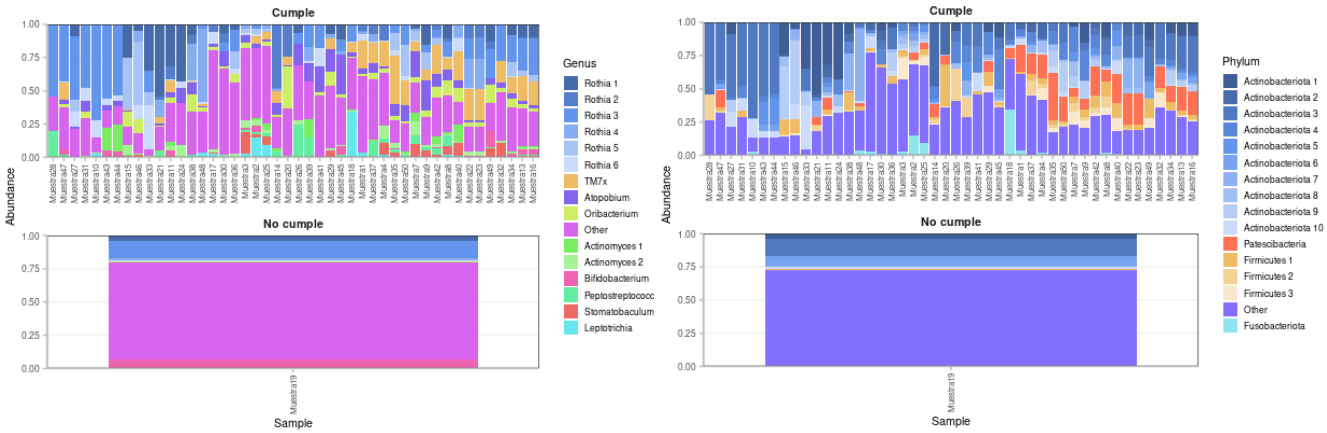

ILUSTRACIÓN 14 ABUNDANCIA RELATIVA RESPECTO A LA CONCENTRACIÓN DE TRABAJO, PREDOMINANDO ROTHIA A NIVEL DE GÉNERO Y ACTINOBACTERIOTA A NIVEL DE FILO.

Pureza de trabajo

Diversidad Alfa de acuerdo con la Pureza de Trabajo

Se analizó la pureza de trabajo la cual se clasificó en tres grupos para su estudio, alto, medio y bajo. En la figura A, 23 muestras corresponden a una pureza de ADN alta, 10 a una pureza media y 14 a una pureza de ADN baja, obteniendo como valor de p 0.63. En el análisis de Simpson se obtuvieron 22 muestras con una pureza de ADN alta, 14 con pureza media y 9 con una pureza baja, obteniendo como valor de p 0.29, no encontrando resultados estadísticamente significativos.

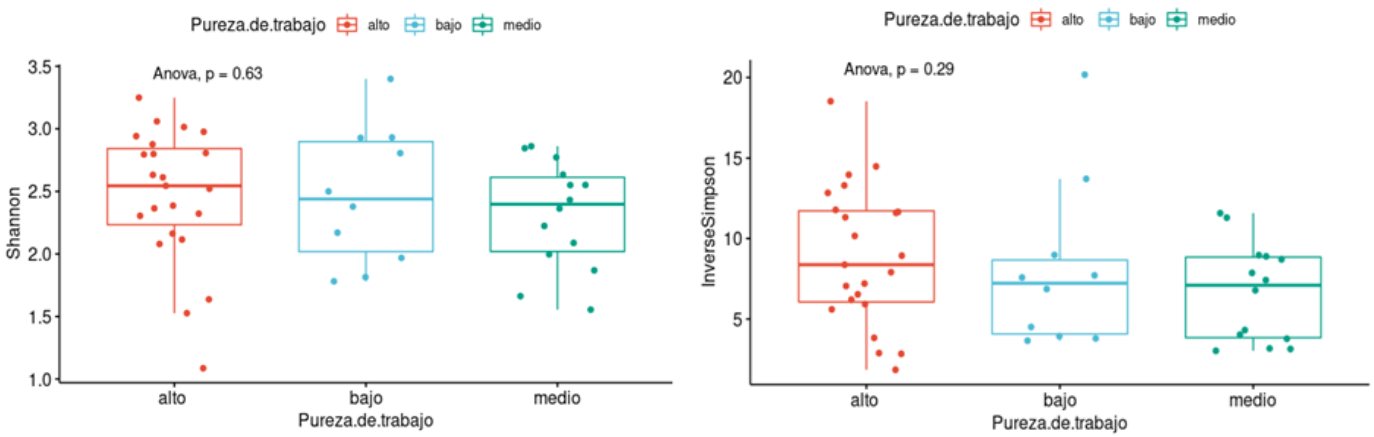

ILUSTRACIÓN 15 DIVERSIDAD ALFA RESPECTO A LA PUREZA DE ADN, CLASIFICÁNDOLA EN ALTA, MEDIA Y BAJA CON UN VALOR DE P=0.53

Diversidad Beta de acuerdo con la Pureza de ADN

Posteriormente se obtuvo la diversidad beta para ver si había una diferencia entre ellas, en la cual no se encontró una aglomeración específica de la misma utilizando los métodos de UniFrac ponderado y UniFrac no ponderado y Brai Curtis esto con la finalidad de poderlo medir cualitativa y cuantitativamente. Sin encontrar diferencias estadísticamente significativas.

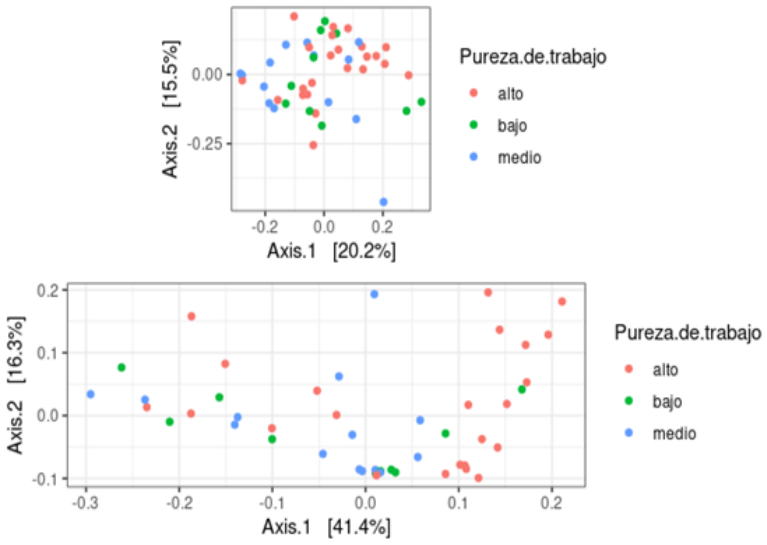

ILUSTRACIÓN 16 DIVERSIDAD BETA RESPECTO A LA PUREZA DE ADN, DONDE SE DESCRIBE CUALITATIVA Y CUANTITATIVAMENTE SIN PRESENTAR AGLOMERACIONES.

Abundancia Relativa respecto a la Pureza de ADN

Se obtuvo la abundancia de la Pureza de ADN clasificándose en tres grupos pureza alta, baja y media, obteniendo para cada una de estas la abundancia a nivel de filo (A) y género (B).

En la figura A en las tres categorías predominó *Actinobacteriota* presentando 10 variantes, con 51.3% en pureza de ADN alta, 66.4% en baja y 69% en media, mientras que a nivel de género *Rothia* fue la más abundante 29.2% en una pureza de ADN alta ,48.8% a una pureza de ADN media y 40.4% correspondiente a una pureza baja, presentó la mayor abundancia mostrando 6 variantes de la misma.

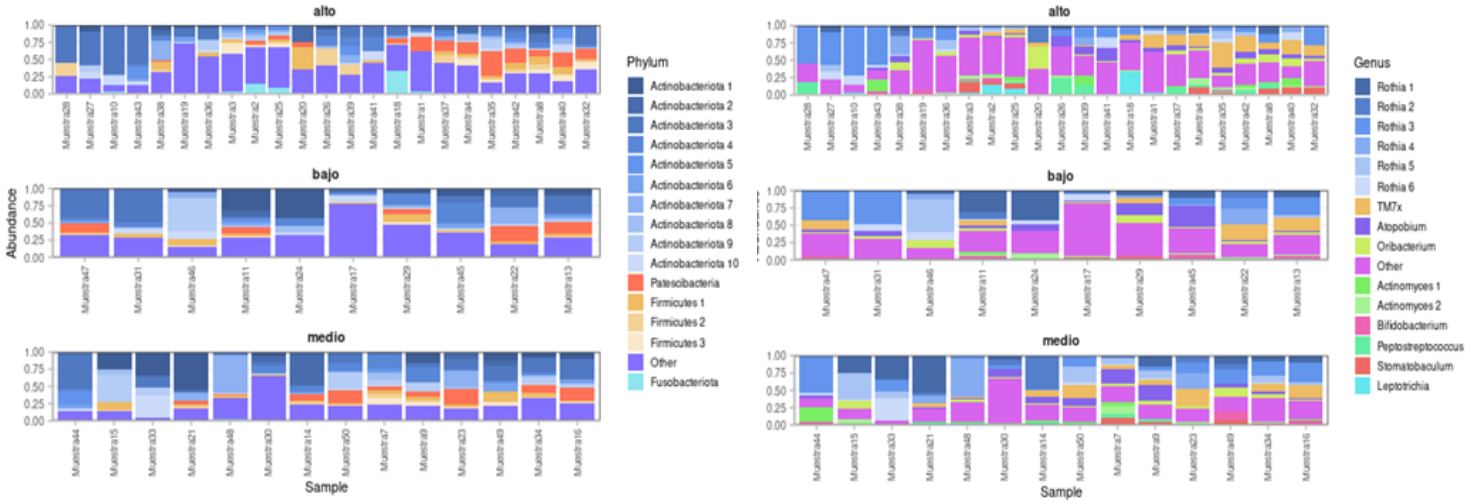

ILUSTRACIÓN 17 ABUNDANCIA RELATIVA RESPECTO A LA PUREZA DE ADN, PREDOMINANDO ROTHIA A NIVEL DE GÉNERO Y ACTINOBACTERIOTA A NIVEL DE FILO.

Para poder tener una mejor visión de que taxas eran propios de cada grupo y cuantas compartían, se obtuvo el diagrama de Venn, el cual demostró que de las 50 muestras solo una de ellas pertenecía a una pureza de ADN baja y una a una pureza media, demostrando que no había diferencias significativas, además de encontrarse 89.9% sin núcleo.

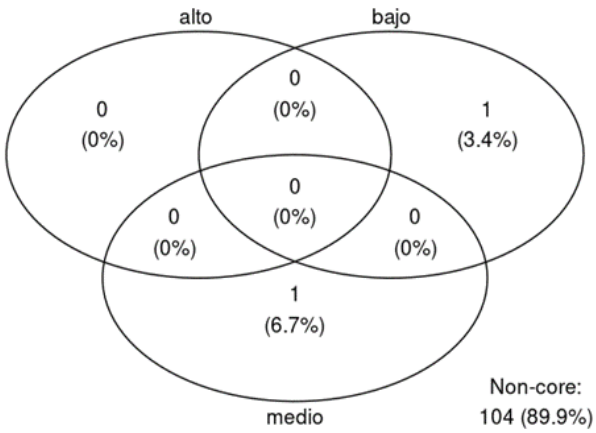

ILUSTRACIÓN 18 DIAGRAMA DE VENN PARA CALIDAD DE MATERIAL GENÉTICO BACTERIANO ALTA O MEDIA, DONDE SOLO UNA MUESTRA CORRESPONDIÓ A CALIDAD MEDIA.

Análisis Diferencial en la Pureza de ADN

Se obtuvo la diferencia entre la pureza del ADN, donde se encontraron aumentados los géneros *Moginacterium, Brucella, Oribacterium, Actinomyces y Candidatus Saccharimonas,* y a nivel de Filo *Patescibacteria, Actinobacteria, Firmicutes y Proteobacteria* se encontraron aumentados de igual manera, encontrando que hay diferencia estadísticamente significativa en las muestras con pureza de alta calidad.

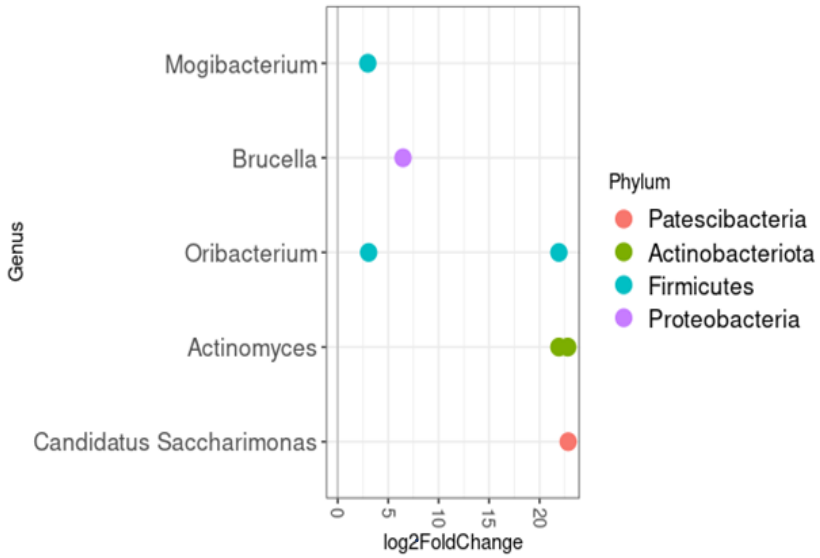

ILUSTRACIÓN 19 EL ANÁLISIS DIFERENCIAL RESPECTO A LA PUREZA DE ADN, MUESTRA FILO Y GÉNERO AUMENTADOS, MOSTRANDO UNA DIFERENCIA ESTADÍSTICAMENTE SIGNIFICATIVA

Diversidad alfa entre el grupo con calidad alta y calidad media

La riqueza y uniformidad en cada muestra se estimaron utilizando las especies observadas y los índices de Shannon y Simpson, estos dos índices son comúnmente utilizados para la descripción cuantitativa y cualitativa de la diversidad alfa en polimorfismos comunitarios. El índice de Shannon está en proporción directa a la diversidad microbiana en las muestras, mientras que el índice de Simpson se correlaciona negativamente con la diversidad del microbiota.

Se analizó la calidad del material genético tomando en cuenta tres criterios, concentración, integridad y pureza del mismo, para así poder categorizarla en calidad alta, media y baja. Una vez realizada la asignación se prosiguió a calcular la diversidad alfa para ver si contaba con alguna diferencia entre ambos grupos. Teniendo como resultado una p=0.67 usando Shannon, mientras que Simpson obtuvo un valor de p=0.96 y así se comprobó que no existe diferencia estadísticamente significativa en ninguno de los dos índices usados.

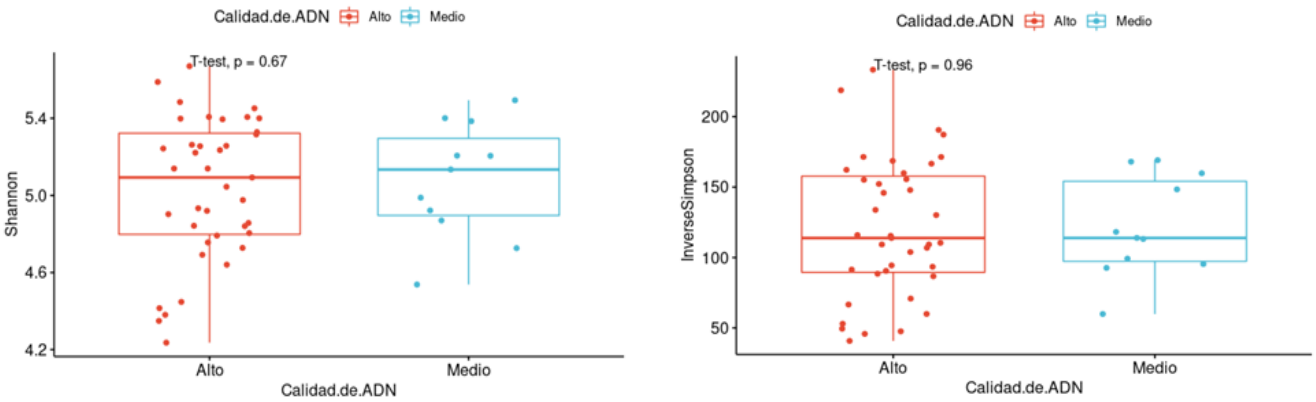

ILUSTRACIÓN 10 DIVERSIDAD ALFA DE LA CALIDAD DE ADN, EN LA CUAL SE ENGLOBO, CONCENTRACIÓN, INTEGRIDAD Y PUREZA DE TRABAJO CON UNA P= 0.67

Diversidad Beta entre el grupo con Calidad alta y Calidad media

Se realizó el análisis de coordenadas principales (PCoA), basado en la distancia UniFrac no ponderada y la distancia UniFrac pondera con la finalidad de observar la existencia de relación de la estructura microbiana de los grupos de manera cuantitativa y cualitativa con la cual no se observó diferencia alguna.

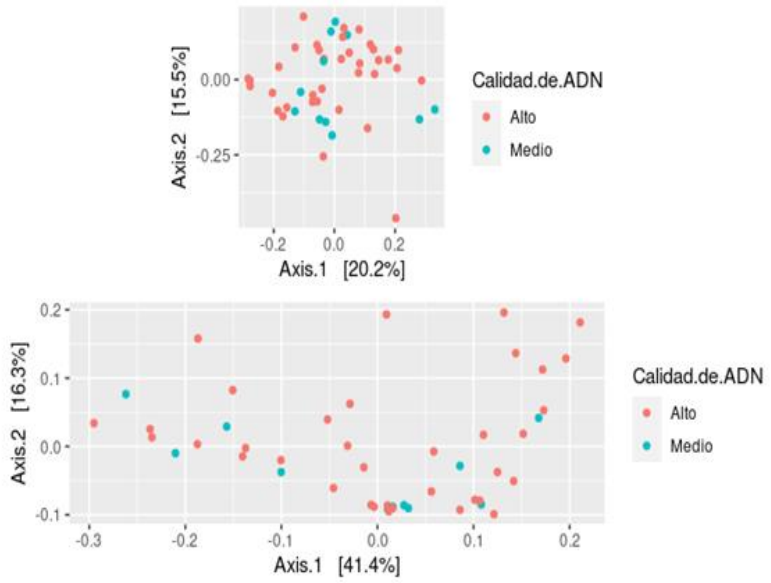

ILUSTRACIÓN 21 DIVERSIDAD BETA DE LA CALIDAD DE ADN, DONDE SE DESCRIBE CUALITATIVA Y CUANTITATIVAMENTE SIN PRESENTAR AGLOMERACIONES

.

Abundancia Relativa conforme a Calidad Alta y Media del Material Genético Bacteriano.

Se obtuvo la abundancia relativa sobre la calidad de ADN. Se muestra en la figura (A) a nivel de Género, donde *Rothia* estuvo presente con una abundancia de 37.1% en aquellas muestras con una calidad alta y presentó 38.5% en calidad media, siendo el filo más abundante, además de expresar seis diferentes tipos de posibles variantes, seguida por *Actinomyces* con 10.6% en calidad alta y 9.2% en media, TM7x obtuvo un 8.3% en calidad alta y 7.4% en calidad media, mientras que a nivel de Filo en la figura (B) podemos ver que *Actinobacteriota* fue la más abundante con

un 58.5% en calidad alta y un 64.1% en calidad media, seguida por *Firmicutes* con 19.9% en calidad alta y 16.2% en calidad media, mientras que *Patescibacteria* presentó 12.4% en calidad alta y 14.4% en media.

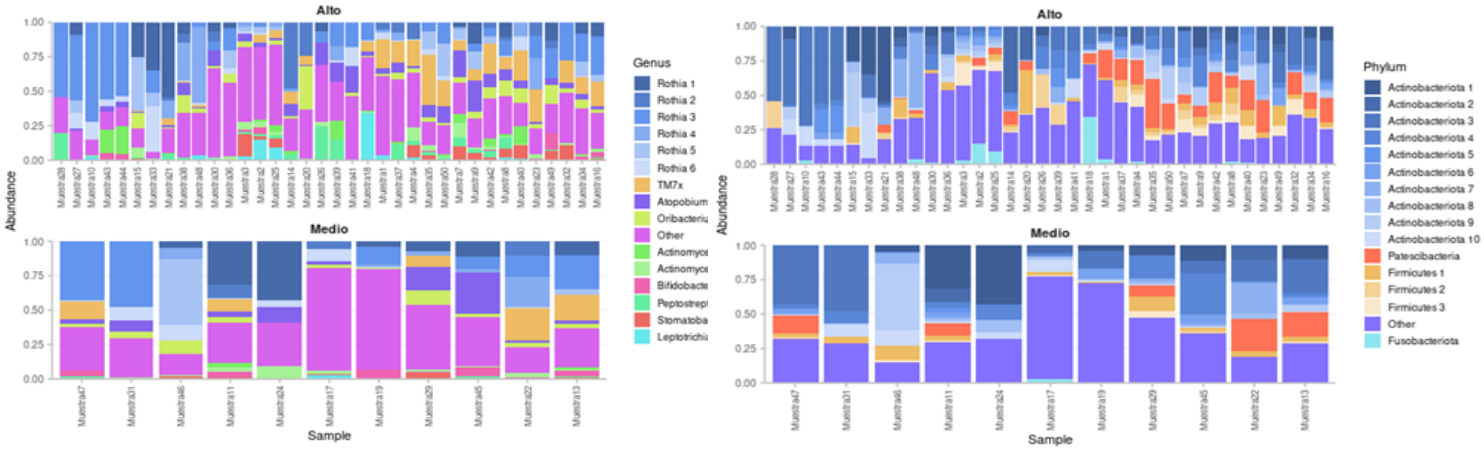

ILUSTRACIÓN 22. ABUNDANCIA RELATIVA CONFORME A CALIDAD ALTA Y MEDIA DE MATERIAL GENÉTICO BACTERIANO PREDOMINANDO ROTHIA A NIVEL DE GÉNERO Y ACTINOBACTERIOTA A NIVEL DE FILO.

Para poder tener una mejor visión de que taxa eran propias de cada grupo y cuantas compartían, se obtuvo el diagrama de Venn en el cual, se encontró que una taxa fue propia de calidad de material genético bacteriano media, equivalente al 3.4%.

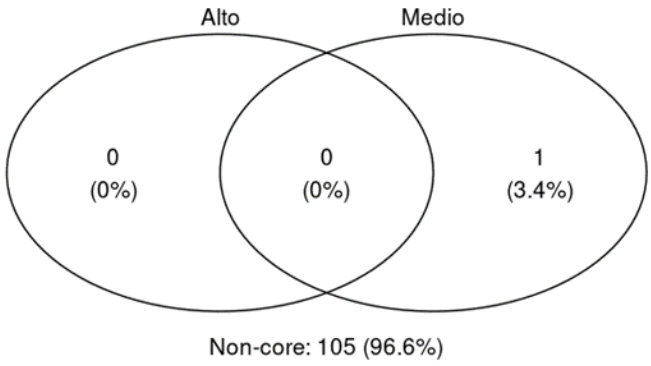

ILUSTRACIÓN 23 DIAGRAMA DE VENN PARA CALIDAD DE MATERIAL GENÉTICO BACTERIANO ALTA O MEDIA, DONDE SOLO UNA MUESTRA CORRESPONDIÓ A CALIDAD MEDIA.

Análisis Diferencial entre el grupo de Calidad Alta contra Calidad Media.

Resultado del análisis, se encontró que a nivel de filo *Actinobacteriota* fue diferencial estando en mayor proporción en el grupo de calidad alta, con una padj=6.34E; mientras que en el género de *Rothia* se encontraron dos variantes, la primera con un valor de padj=10.71 y la segunda con un valor de padj=10.32.

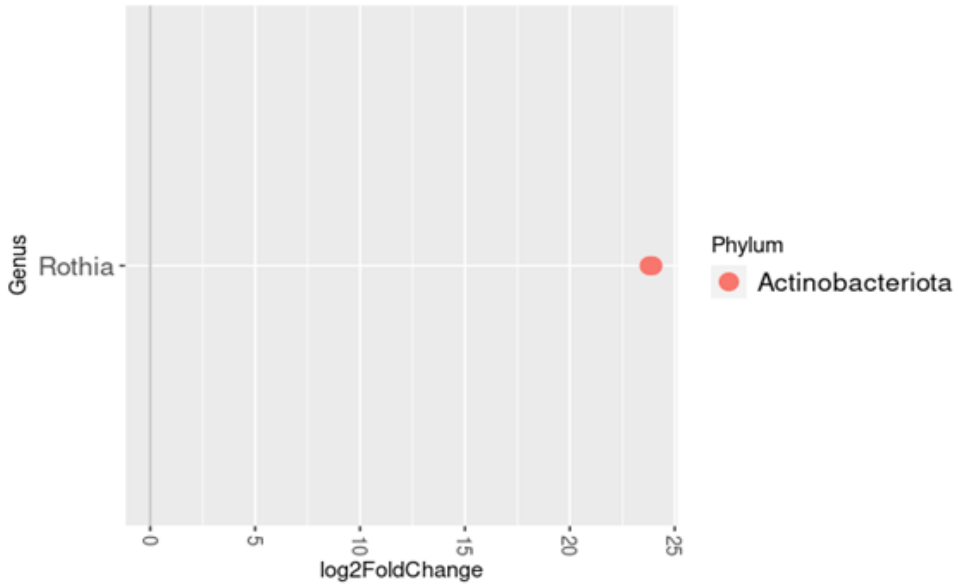

ILUSTRACIÓN 24 EL ANÁLISIS DIFERENCIAL MUESTRA UNA MAYOR PROPORCIÓN EN CALIDAD ALTA Y LA PRESENCIA DE DOS VARIANTES EN EL GÉNERO

.

TABLA 3 DIFERENCIAS TAXONÓMICAS DE VARIABLES

Variable		Diversidad Alfa		*Filo*	*Género*	*Especie*	log2Fold Change	padj
		Shannon	Simpson					
Año de recolección	Antes 2015= 39	0.72	0.36	*Patescibacteria*	*Candidatus Saccharimonas*	NA	22.84	8.11-19
Antes 2015-del 2015 en adelante	2015 en adelante= 11			*Actinobacteriota*	*Actinomyces*	NA(2)	22.79	5.70-12
				Firmicutes	*Oribacterium*	NA(2)	21.91	7.22-10
					Mogibacterium	NA	2.98	0.47
				Proteobacteria	*Brucella*	NA	6.47	0.47
Kit de extracción	QIAGEN= 19	0.86	0.49	*Firmicutes*	*Oribacterium*	NA	-23.14	9.27-21
QIAGEN-Mobio	Mobio= 31			*Patescibacteria*	*NA*	NA	-22.56	9.15-3
				Proteobacteria	*Brucella*	NA	-7.44	1.54-4
Pureza de trabajo	Alta= 25	0.63	0.29	*Patescibacteria*	*Candidatus Saccharimonas*	NA	22,84	8.11-05
Alta-baja	Baja= 10			*Actinobacteriota*	*Actinomyces*	NA(2)	22,79	5.70-12
				Firmicutes	*Oribacterium*	NA(2)	21,91	7.76-10
					Mogibacterium	NA	2,98	0,047
				Proteobacteria	*Brucella*	NA	6,47	0,047
Pureza de trabajo	Alta= 25	0.63	0.29	*Patescibacteria*	*Candidatus Saccharimonas*	NA	22,84	8.11-05
Alta-media	Media=15			*Actinobacteriota*	*Actinomyces*	NA(2)	22,79	5.70-12
				Firmicutes	*Oribacterium*	NA(2)	21,91	7.76-10
					Mogibacterium	NA	2,98	0,047
				Proteobacteria	*Brucella*	NA	6,47	0,047
Pureza de trabajo	Media=15	0.63	0.29	*Fusobacteriota*	*Leptotrichia*	NA	-19,21	4.58-10
Media-baja	Baja= 10			*Firmicutes*	*Oribacterium*	NA	-20,98	2.95-7
Calidad de ADN	Alta= 39	0.67	0.96	*Actinobacteriota*	*Rothia*	NA(2)	23.81	6.24-24
Alta-media	Media= 11							

9.0 Discusión

En el presente estudio se realizó el análisis de secuencias pertenecientes a 50 muestras del Laboratorio de Genómica Clínica de la División de Posgrado de la Facultad de Odontología, UNAM; Las cuales contaban con datos de recolección (conforme a las buenas prácticas), almacenamiento (tiempo y características del mismo) y procesamiento de las mismas (concentración, integridad y pureza). Además del análisis primario de secuenciación.

Referente al análisis primario de los datos, los resultados fueron de acuerdo con los parámetros de referencia sugeridas propiamente por Ilumina, los cuales se encuentran alojados en la página de la misma.

En cuanto a la recolección de las muestras se encontró que contaron con las mismas características (tiempo de recolección, el evitar lavar la cavidad bucal, ingerir algún tipo de alimentos y que se recolectará solo saliva basal) establecidas por Wong y col, Ribar y col y Poussin y col, (11, 14, 15)

Con respecto a las condiciones de almacenamiento y preservación de la muestra, los métodos de preservación fueron a -80 ° C, siendo similar a lo establecido por Rob Knight y col. y Salter y col, lo cual es consistente con los estudios publicados previamente en donde al evaluar la riqueza en función de OTU´s no difirió significativamente entre las condiciones de almacenamiento.

Respecto al tiempo de almacenamiento se encontró aumento a nivel de Filo de *Actinobacteriota,* y *Firmicutes* lo cual es similar a lo encontrado por Choo y col. (2015) (16) donde compararon seis diferentes condiciones de almacenamiento de muestra; En la misma investigación ellos reportan la presencia de *Bacteroidetes*, los cuales no se expresaron en este estudio. (17)

Conforme a la calidad de librerías se encontró a nivel de Filo una mayor abundancia de *Actinobacteriota*, seguido por *Actinobacteria* y *TM7x*, también se encontró

presencia de *Firmicutes*, lo cual es consistente con Lazarevic y col, donde reportan la presencia *Firmicutes*, *Proteobacteria*, *Actinobacteria*, *Fusobacteria*, *TM7x* y *Spirochaetes.* Solo nuestra muestra 11 no contó con una calidad de librerías adecuada, lo cual se puede deber a que el paciente que donó la muestra trabajaba constantemente con sustancias tóxicas y así como lo menciona Li y col, el ambiente y la exposición a químicos puede ser un factor para llevar a cabo una modificación de la microbiota. (18, 19)

Respecto a la calidad de las lecturas y conforme al lote de kit en donde se analizaron las diferencias entre Mobio BS16C9 y Qiagen 157042821 en los cuales no se encontraron diferencias significativas, lo cual es similar con estudios previos de Mckenzie y col, sin embargo, encontramos en el kit Mobio BS16C9 disminuidas a nivel de Filo a *Patescibacteria* y *Proteobacteria* y a nivel de género *a Brucella.* En Qiagen 157042821 se encontró disminuida a nivel de Filo *Firmicutes* y en Género a *Orinobacterium.* Lo cual concuerda con los estudios publicados previamente en donde se encontraron disminuidas bacterias gram-positivas (*Firmicutes y Actinobacteria*) tal como lo reportan Mckenzie y col. Lo que sugiere que este kit puede no ser lo suficientemente estricto para la lisis óptima de algunos organismos Gram-positivos. (20, 21)

Para analizar la calidad de ADN de las 50 muestras se utilizaron tres criterios el primero de ellos fue la Integridad del ADN bacteriano en el cual no encontramos ADN degradado en ninguna de las muestras, a nivel de Filo predominó *Actinobacteriota* y a nivel de Género *Rothia* fue la más abundante, lo cual difiere con los resultados de Alhaddad y col, ya que ellos encontraron ADN degradado al momento de realizar la electroforesis en gel con grandes extensiones del mismo sin ADN genómico según lo reportan, lo cual puede estar relacionado con el tamaño de la muestra. (22)

El segundo criterio fue la concentración de ADN bacteriano en donde solo la muestra 19 no cumplió con los parámetros indicados, siendo baja, al igual que Kuchler y col y Alhaddad y col, la concentración se determinó por espectrofotometría

usando NanoDrop al igual que en este estudio, donde ellos obtuvieron amplios rangos de cantidades de ADN. (22, 23)

En lo que concierne a la pureza del material genético en la cual se obtuvo un valor de p=0.302 lo cual no es estadísticamente significativo siendo consistente con Wagner y col, en donde obtuvieron un valor de p=0.3 bajo los parámetros de 260/280 y utilizando el kit de extracción Mobio al igual que nosotros. (21)

En cuanto a la calidad del ADN, en dicho estudio se decidió contar con tres valores (integridad, pureza y concentración), para así contar con una calidad global del material genético; al llevar cabo la revisión de la literatura Alhhadad y col, tomaron en cuenta los mismos tres criterios para determinar la calidad de sus muestras, concordando con nuestros resultados donde no obtuvimos diferencias estadísticamente significativas al igual que ellos, la gran mayoría de las muestras extraídas en este estudio se encontraban dentro del rango aceptado de pureza de ADN y en general, se obtuvo una alta calidad de ADN genómico en forma de una banda única visible. Ellos presentaron una mayor inconsistencia en sus resultados, como degradación de ADN, baja pureza y variabilidad en la concentración de ADN de algunas muestras, creemos que esto puede ser debido al tamaño de las muestras y la cantidad de células salivales que se utilizaron para este estudio. (22)

10.0 Conclusiones

Nuestro estudio destaca varios aspectos importantes al analizar cómo influye la recolección de la muestra, el tiempo de almacenamiento y preservación de estas. Las condiciones de almacenamiento podrían tener el potencial de introducir alteraciones sustanciales en el perfil de la comunidad microbiana basado en la secuenciación del gen 16S rRNA, cuando no se siguen los protocolos, buenas prácticas y procedimientos estandarizados, tales como congelación a -80 °C. Una limitación del estudio son los ciclos de congelación y descongelación de las muestras. Además de mejorar los procedimientos estandarizados al momento de la recolección de la muestra.

Este estudio demuestra que al analizar la calidad de las lecturas en función del lote de kit de extracción de ADN bacteriano podemos concluir que no hubo diferencias relevantes entre alguno de los dos kits que se utilizaron, además de demostrar que es un paso clave en los estudios de metataxonómica, donde una posible limitación es que puede no ser lo suficientemente estricto para la lisis óptima de algunos organismos Gram-positivos.

En general, la calidad de las lecturas con base en la integridad, pureza y concentración de ADN bacteriano fue adecuada y estuvo dentro de los parámetros establecidos, demostrando que las muestras tuvieron una buena calidad de ADN bacteriano, sin embargo, tuvo diferencias respecto a la revisión bibliográfica donde la N del estudio pudo influir en estos resultados.

Actualmente la literatura relacionada con dicho tema es escasa por lo cual el realizar buenas prácticas de análisis para estudios de metataxonómica sería un gran avance en el tema, evitando sesgos y ser completamente reproducibles.

11. 0 Referencias

1. Zhang Y. Human oral microbiota and its modulation for oral health. In: Wang X, editor. Elsevier ed. Biomedicine and Pharmacotherapy: Science direct; 2018. p. 883-93.
2. Seesandra RV. The human microbiome and cancer. Cancer Prevention Research [Internet]. 2017; 10(4). Available from: https://cancerpreventionresearch.aacrjournals.org/content/10/4/226.
3. Consortium HMP. A framework for human microbiome research. Nature. 2012;486(7402):215-21.
4. Ritchie Al. Metagenomic Characterization of the Respiratory Microbiome. A Piece de Resistance. American Thoracic Society [Internet]. 2020; 202(feauturing article). Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7397782/.
5. Julian MR. The vocabulary of microbiome research: a proposal2015. Available from: https://microbiomejournal.biomedcentral.com/articles/10.1186/s40168-015-0094-5#rightslink.
6. Sarah HK. Metataxonomic and Metagenomic Approaches vs Culture Based Techniques for Clinical Pathology. Microbial frontal [Internet]. 2016. Available from: https://pubmed.ncbi.nlm.nih.gov/27092134/.
7. Fabiola VG. The 16S rRNA gene in the study of marina microbial communities. 2015:297-313.
8. Case RJ, Boucher Y, Dahllof I, Holmstrom C, Doolittle WF, Kjelleberg S. Use of 16S rRNA and rpoB genes as molecular markers for microbial ecology studies. Applied and Environmental Microbiology. 2007;73(1):278-88.
9. Sinha R, Abnet CC, White O, Knight R, Huttenhower C. The microbiome quality control project: baseline study design and future directions. Genome Biol. 2015;16:276.
10. von Wintzingerode F, Göbel UB, Stackebrandt E. Determination of microbial diversity in environmental samples: pitfalls of PCR-based rRNA analysis. FEMS Microbiol Rev. 1997;21(3):213-29.
11. Poussin C, Sierro N, Boué S, Battey J, Scotti E, Belcastro V, et al. Interrogating the microbiome: experimental and computational considerations in support of study reproducibility. Drug Discov Today. 2018;23(9):1644-57.
12. Santella RM. Approaches to DNA/RNA Extraction and whole genome amplification. Cancer Epidemiol Biomarkers Prev. 2006;15(9):1585-7.
13. Salter SJ, Cox MJ, Turek EM, Calus ST, Cookson WO, Moffatt MF, et al. Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. BMC Biol. 2014;12:87.
14. Knight R, Vrbanac A, Taylor BC, Aksenov A, Callewaert C, Debelius J, et al. Best practices for analysing microbiomes. Nat Rev Microbiol. 2018;16(7):410-22.
15. Bruinsma FJ, Joo JE, Wong EM, Giles GG, Southey MC. The utility of DNA extracted from saliva for genome-wide molecular research platforms. BMC Res Notes. 2018;11(1):8.
16. Choo JM. Sample storage conditions significantly influence faecal microbiome profiles 2015. Available from: https://www.nature.com/articles/srep16350.
17. Bahl MI, Bergström A, Licht TR. Freezing fecal samples prior to DNA extraction affects the Firmicutes to Bacteroidetes ratio determined by downstream quantitative PCR analysis. FEMS Microbiol Lett. 2012;329(2):193-7.
18. Lazarevic V, Whiteson K, Huse S, Hernandez D, Farinelli L, Osterås M, et al. Metagenomic study of the oral microbiota by Illumina high-throughput sequencing. J Microbiol Methods. 2009;79(3):266-71.

19. Li N, Li J, Zhang Q, Gao S, Quan X, Liu P, et al. Effects of endocrine disrupting chemicals in host health: Three-way interactions between environmental exposure, host phenotypic responses, and gut microbiota. Environ Pollut. 2021;271:116387.
20. Kennedy NA, Walker AW, Berry SH, Duncan SH, Farquarson FM, Louis P, et al. The impact of different DNA extraction kits and laboratories upon the assessment of human gut microbiota composition by 16S rRNA gene sequencing. PLoS One. 2014;9(2):e88982.
21. Wagner Mackenzie B, Waite DW, Taylor MW. Evaluating variation in human gut microbiota profiles due to DNA extraction method and inter-subject differences. Front Microbiol. 2015;6:130.
22. Alhaddad H, Maraqa T, Alabdulghafour S, Alaskar H, Alaqeely R, Almathen F, et al. Calidad y cantidad de muestras de ADN de camello dromedario de sangre total, saliva y pelo de la cola. PLoS One. 2019;14(1):e0211743.
23. Küchler EC, Tannure PN, Falagan-Lotsch P, Lopes TS, Granjeiro JM, Amorim LM. Buccal cells DNA extraction to obtain high quality human genomic DNA suitable for polymorphism genotyping by PCR-RFLP and Real-Time PCR. J Appl Oral Sci. 2012;20(4):467-71.