UNIVERSIDAD NACIONAL DE LA MATANZA
DEPARTAMENTO DE INGENIERÍA E INVESTIGACIONES TECNOLÓGICAS

PROGRAMA PROINCE
CÓDIGO C169

APLICACIONES DE DATA MINING AL ESTUDIO DEL MICROBIOMA HUMANO

Director: Santa María, Cristóbal Raúl
Co-director: López, Luis
Integrantes: Soria, Marcelo Abel; Martínez, Pablo Witold; Otaegui, Juan Carlos; Santa María, Victoria; Galanternik, Fernando; Ávila, Laura

Fecha de Inicio: 2015/01/01
Fecha de Finalización: 2016/12/31
Informe Final

1. Resumen y Palabras Clave

Resumen: El proyecto desarrollado procuró aportar procedimientos computacionales adecuados para analizar la relación clínica entre el microbioma intestinal y la presencia de patologías tales como el cáncer de colon y la enfermedad de Crohn. Utilizando un sistema operativo biolinux y el software Superfocus se trabajó con conjuntos de datos formados inicialmente por las secuencias de ADN obtenidas de microbiomas de pacientes. Las muestras fueron extraídas de repositorios metagenómicos internacionales. Mediante los procesos realizados se obtuvieron distribuciones de abundancia de carácter taxonómico y funcional. A partir de ellas se aplicaron métodos para construir la matriz de distancia entre secuencias. Para esta tarea se utilizó la distancia de Jensen-Shannon entre distribuciones. Usando la matriz de distancias se armaron clusters mediante la aplicación del algoritmo PAM y se evaluó la consistencia de los agrupamientos obtenidos. De tal forma se estableció una “pipeline” de procedimientos, programados en lenguaje R, para llegar a agrupamientos de microbiomas que representaran enterotipos clínicos. Para muestras relacionadas con la Enfermedad de Crohn a partir de la distribución de abundancia taxonómica de los microbiomas se analizó la relación entre el diagnóstico clínico y los clusters obtenidos aplicando el método de componentes principales. El trabajo consolidó el conocimiento necesario para afrontar próximos estudios a partir de muestras de pacientes locales en la continuidad de la línea de investigación.
Palabras Clave: Microbioma Cluster Enterotipo Diagnóstico

2. Memoria descriptiva

Introducción

Desde los comienzos de su vida el cuerpo humano es colonizado por bacterias, arqueas, hongos y virus. Esta comunidad de microorganismos se denomina microbioma y contiene diez veces más células que las del propio cuerpo humano. Hay más de 1000 especies de microbios que viven en el intestino humano y conforman el microbioma. Este juega un rol importante en la protección contra patógenos, modula la inmunidad, regula procesos metabólicos y es incluso considerado en algunos casos como un "órgano endocrino". Sin embargo, los métodos tradicionales de cultivo no pueden identificar todas estas especies, por lo que para identificarlos se utilizan métodos de secuenciación de biología molecular [1]. La cantidad de genes presentes en total es varios órdenes de magnitud mayor que la del genoma humano. La nueva generación de tecnologías de secuenciación de ADN ha permitido comenzar a estudiar las características del microbioma humano según la edad de la persona, la localización geográfica, los hábitos alimentarios y la presencia de enfermedades. El objetivo principal de estos estudios metagenómicos es analizar la estructura y la dinámica de comunidades microbianas, para establecer cómo se relacionan sus miembros entre sí, cuáles son las sustancias que producen y que consumen, y especialmente cuáles son sus interacciones con las células humanas próximas y cómo se modifica la comunidad en presencia de enfermedades. El estudio por medio de la identificación de un gen marcador como el 16s rRNA pretende evaluar características ecológicas como la riqueza y la diversidad, mientras que el análisis del metagenoma como un todo, identificando las secuencias obtenidas por comparación con una base de datos genética previamente armada, permite agrupar los genes por funciones metabólicas asociadas con la presencia de enfermedades. Diferentes experiencias prueban que estos estudios sobre todo el metagenoma producen mejores resultados para la apreciación clínica de diferentes patologías que los que arroja el trabajo con genes marcadores [2]. La secuenciación del metagenoma permite determinar la potencialidad funcional codificada dentro del microbioma y ayuda a entender la interacción entre los patógenos y el organismo humano. En tal sentido puede ser útil para el diagnóstico y prevención de enfermedades. Sin embargo, en el estudio del metagenoma del microbioma ha sido reportado, en múltiples casos, contaminación del ADN huésped por lo que el desarrollo de mejores métodos experimentales es mandatorio para arribar a conclusiones [3].

Los estudios que emplean tales técnicas se multiplican velozmente y existen a nivel mundial proyectos de investigación como el Metagenomics of the Human Intestinal Tract (MetaHIT) o el Human Microbiome Proyect (HMP). En nuestro país en el marco del Plan Nacional de Ciencia, Tecnología e Innovación (Argentina Innovadora 2020), dentro del área de Salud, se ha comenzado a desarrollar una Plataforma Tecnológica de Genómica y Bioinformática que facilitará estudios similares. El presente informe corresponde al proyecto que se ha desarrollado en el período 2015-2016 en el marco del Programa de Incentivos. Se prosiguió así con el objetivo de entender el funcionamiento del microbioma humano a partir del procesamiento y análisis de muestras de secuencias de ADN, y de desarrollar nuevas herramientas para analizar y caracterizar el curso de patologías poniendo énfasis en el cáncer de colon y en la enfermedad de Crohn.

De acuerdo a lo expuesto hasta aquí el inicio del trabajo computacional consiste en obtener los datos secuenciados de una muestra integrada por varios microbiomas. Para ello existen dos alternativas: buscar datos en repositorios internacionales u obtenerlos por vía de estudios sobre pacientes locales. Por razones operativas y

presupuestarias se decidió cumplimentar esta etapa del trabajo tomando datos del repositorio de NCBI (National Center of Biotechnology Information). A su vez se decidió también estudiar en primer término las posibilidades del análisis utilizando un gen marcador. De tal modo, cada microbioma de la muestra fue cotejado con una base de datos correspondiente a dicho gen para encontrar la distribución de frecuencias de los microorganismos identificados por él. Alternativamente el conjunto de secuencias del microbioma puede compararse con otra base de datos de funciones genéticas para agrupar los genes integrantes por función y así obtener la distribución de frecuencias según las funciones metabólicas que las secuencias integrantes revelan [4]. Sin embargo la tecnología disponible, en forma libre, para esta tarea presenta aun deficiencias e incompatibilidades con los sistemas operativos instalados (Linux y Biolinux) razón por la cual luego de reiterados intentos se optó por continuar la línea de trabajo con genes marcadores al menos en esta etapa.

En cualquier caso una vez formadas las matrices que representan por fila las distribuciones de cada microbioma individual estos pueden agruparse en clusters. Los conjuntos obtenidos clasifican a los individuos según características clínicas cuyo valor debe ser sopesado en los aspectos médicos del trabajo. Si tal clasificación tiene significancia clínica en el diagnóstico o pronóstico de las enfermedades analizadas, puede utilizarse con métodos predictivos tales como árboles de decisión para establecer diagnósticos o pronósticos en pacientes aún no clasificados.

Se han estudiado en una primera aproximación los métodos computacionales más adecuados para estas tareas [5] y de acuerdo a ello se los ha utilizado aplicando programas desarrollados en R [6] y software Infostat.[7]

La metagenómica tiene sus límites dado que describe solamente las funciones potenciales de la bacteria y no su actividad real. Para el estudio de la funcionalidad de las bacterias se utiliza el metatrasncriptoma, que es la secuenciación de ARNm y permite hacer un perfil funcional del microbioma bajo diferentes circunstancias. Sin embargo, el estudio de ARNm tiene muchas limitaciones y es muy dificultoso por lo que pocos estudios se han realizado hasta la fecha.

Materiales y Métodos

Se utilizaron dos conjuntos de datos. El primer conjunto corresponde a las Distribuciones de Abundancia de Microorganismos de acuerdo al gen marcador 16S rRNA en los microbiomas de 22 pacientes de origen europeo, 13 japoneses y 4 norteamericanos. Las secuencias de ADN originales fueron obtenidas por el método Sanger. Estos datos fueron utilizados para rehacer el camino de análisis llevado a cabo por Arumugan et al. en [8], de forma de lograr conocimiento y práctica en el método de trabajo. El segundo conjunto fue extraído de la base BioProyect de NCBI donde figura bajo el número de acceso PRJNA 175224. El conjunto fue obtenido por secuenciación Illumina con dos lecturas por secuencia y corresponde a las muestras de 11 pacientes, 7 de los cuales están sanos y los otros cuatro tienen la enfermedad de Crohn [9]

Con el segundo conjunto hubo que desarrollar toda la tarea de construir las Distribuciones de Abundancia de los 11 microbiomas a partir de las secuencias obtenidas en formato fastq. Un primer aspecto a resolver fue el del hardware y sistema operativo necesario para el trabajo. Con vistas a la línea de investigación que el proyecto abrió dentro del Departamento de Ingeniería e Investigaciones Tecnológicas de la UNLAM se decidió adaptar un servidor para el trabajo del grupo, ya que el volumen de memoria y la capacidad de proceso requeridas por los conjuntos de datos utilizados en la bibliografía y

los que potencialmente pudieran formarse luego a partir de muestras propias, lo hacían necesario. También se decidió dotar a este servidor con el sistema BioLinux, de uso habitual en bioinformática pues posee una fácil interacción con paquetes de software libre, eficientes y probados, en la investigación en biología computacional. Además de las tareas de instalación del conjunto, esto implicó la dedicación al aprendizaje de la herramienta por parte de integrantes del grupo con los consiguientes ensayos de prueba y error. Además en ocasiones que se detallan más adelante se han utilizado máquinas virtuales armadas en la nube Amazon para llevar adelante las tareas.

Para cobrar idea del volumen de datos a utilizar se comenzó estableciendo que espacio promedio, calculado en bytes, ocupa una secuencia. Al respecto el código ASCII utiliza 7 bits para representar caracteres más un bit de paridad. Cada letra que representa a una base es entonces un carácter de escritura que ocupa 1 byte (es decir 8 bits). En términos electrónicos cada bit puede estar "prendido" o no lo que en numeración binaria se lee como 1 o 0 respectivamente. Así las cosas en la memoria de una computadora las letras que representan a las bases tienen la siguiente expresión en código ASCII:

C 0110 0011 citosina
G 0110 0111 guanina
A 0110 0001 adenina
T 0111 0100 timina

En Sistema Internacional Decimal 1 K es un kilobyte o sea 1000 bytes=10^3 (si se usa base 2 1K son 1024 bytes= 2^{10}). Entonces un megabyte Mb=1000000=10^6. Los relevamientos de ADN total de microbiomas producen fragmentos cortos de ADN. Dependiendo de la tecnología de secuenciación y la configuración de los equipos se puede controlar el tamaño y cantidad de estos fragmentos. Con los equipos MiSeq y HiSeq de Illumina, que es la más usada hoy en día, las secuencias de ADN pueden tener entre 150 y 300 bases de largo y la cantidad puede llegar a varios millones de secuencias por corrida. En un trabajo típico de análisis de microbiomas es común tener unos 30 millones de secuencias por cada muestra analizada. Esto determina que sean críticos los tiempos de ejecución de los programas utilizados para el análisis. Hay que aclarar que existen diferentes tipos de análisis que se pueden llevar a cabo con estudios de secuenciación metagenómica, por ejemplo, obtener "contigs", que son ensambles de secuencias en un fragmento bastante mayor y que permite la reconstrucción de genes completos, genomas parciales o, incluso, en algunos casos genomas completos. Otras aproximaciones posibles son la asignación taxonómica de las secuencias, es decir, determinar con la mayor precisión posible a que especie, género, familia, etc. de microorganismos pertenecen las secuencias obtenidas; o determinar una asignación funcional, que consiste en encontrar para las secuencias que codifican proteínas, qué tipo de proteínas y en qué actividades celulares participan.

En este proyecto se pretendieron usar asignaciones taxonómicas y funcionales [4]. Para esto se realizaron pruebas con un software de reciente desarrollo, SUPERFOCUS,[10] que, en principio, debiera efectuar ambas tareas, la determinación taxonómica y la asignación funcional. Para esta última actividad el software utiliza la base de datos SEED [11] que en primer lugar asigna una función a una proteína, luego estas funciones se corresponden a categorías funcionales más generales, denominadas subsistemas, o más específicamente subsistemas de nivel 3. Estos subsistemas a su vez se agrupan en subsistemas más generales de nivel 2, y finalmente se agrupan en la categoría más general que es la del subsistema de nivel1.

SUPERFOCUS corre en plataformas Linux y está diseñado para aprovechar las características multiprocesador de las máquinas modernas. Sin embargo, con archivos de 30 millones de secuencias, como los mencionados antes, la ejecución se vuelve errática,

debido a que el programa no aborta y permanece en ejecución pero sin completar las tareas, incluso por días. Las pruebas iniciales se realizaron en el servidor BioLinux con ocho procesadores Intel-I7 y 16 GB de memoria RAM. Se continuaron las pruebas en el servicio de computación en la nube con una máquina virtual Linux del tipo “c3.8xlarge” (32 procesadores Intel Xeon E5-2680 v2 y 60 GB de memoria), pero se siguieron presentando los mismos problemas.

A continuación se realizaron pruebas con datasets de tamaño creciente para determinar cuál era el tamaño máximo de archivo que SUPERFOCUS podía procesar en un tiempo razonable. Se determinó que con archivos de 500,000 secuencias el proceso se completaba con éxito en todas las pruebas, requiriendo entre 25 y 40 minutos en la máquina local. Una característica no documentada es que para cada proceso de 500,000 secuencias se generan una serie de archivos auxiliares de aproximadamente 2 Gb, que no son eliminados al terminar el proceso. De esta forma con varias muestras de 30 millones de secuencias se satura la capacidad del disco rígido en unas pocas corridas. A partir de esta información se diseñó un script de comandos Unix que lee cada archivo a procesar, lo particiona en bloques de 500,000 secuencias cada uno que envía a SUPERFOCUS y, una vez completado el proceso, elimina los archivos auxiliares y procede con el siguiente bloque. De esta manera, a efecto de prueba, en aproximadamente diez días fue posible procesar secuencias de cinco muestras diferentes, cada una con unos 26 millones de secuencias.

Las salidas de SUPERFOCUS son varios archivos de texto, uno con las asignaciones funcionales, tres con la información de subsistemas y otro más con la información taxonómica. Para la información taxonómica se producen una serie de tablas cuyas filas son las asignaciones taxonómicas y cada microbioma en una columna. Estas tablas se presentan para cada una de las categorías taxonómicas usadas en biología: reino, phylum, clase, orden, familia, género y especie. En el caso de la información funcional, la salida está constituida por tablas para la función específica y tres más para cada uno de los subsistemas que define SEED. El procedimiento experimentado permite entonces conocer la distribución de frecuencias de microorganismos o funciones metabólicas de cada paciente. Sin embargo se presentaron problemas de estabilidad en el funcionamiento de SUPERFOCUS ante actualizaciones y cambios en Linux, debe considerarse la discontinuación del Biolinux durante el desarrollo del trabajo. Así surgieron nuevos inconvenientes cuando se decidió procesar el conjunto de datos referidos a enfermedad de Crohn. Las secuencias del microbioma de cada paciente se recogieron de SRA-NCBI en formato FAST-Q pero, a pesar que en las pruebas realizadas con anterioridad este formato no había ocasionado problemas, en este nuevo caso sí lo hizo. En algunos foros otros grupos radicados en distintos países relataban la aparición del mismo problema razón por la cual se decidió convertir los archivos a formato FASTA. Al procesarlos con SUPERFOCUS tampoco se obtuvieron las salidas previstas si no archivos incompletos de información general sin las tablas de abundancia por funcionalidad y un archivo de texto con la información de pantalla referida a la abundancia taxonómica. Este último archivo tuvo que ser procesado con un programa que se desarrolló en lenguaje C para ser ordenado y expresado como .csv a fin de que la distribución de frecuencias por taxonomía pudiera ser leída por los diferentes softwares estadísticos que se utilizarían.

El trabajo realizado a partir de distribuciones de frecuencias establecidas por taxonomía tuvo dos etapas. En la primera se utilizaron los mismos datos usados en [8] a efecto de rehacer el camino de procesamiento allí realizado lo que permitió conocer en detalle los algoritmos aplicados, la teoría que los sustenta y la sensibilidad ante la variación paramétrica. El conjunto fue extraído de [9] y leído por medio de un script R, lenguaje en el que luego se programaron o invocaron los distintos algoritmos. Siguiendo la

metodología citada en [8] se continuó realizando la clasificación en clusters. Para medir la distancia entre dos microbiomas cuyas distribuciones de especies o géneros, por ejemplo, son conocidas se consideró que cada paciente está representado por una distribución de frecuencias estadísticas. De acuerdo a esto hay que medir si dadas dos distribuciones éstas se parecen mucho, poco o nada, a efecto de terminar ubicándolas en clusters. Para ello el artículo [8] cita la distancia de Jensen-Shanon Una muy buena explicación de cómo funciona esta distancia está dada en el paper [12] pero se requiere explicitar antes el concepto de código de mínima redundancia. En [13] se define que dado un ensamble de un número finito de N mensajes y dado un número de dígitos disponibles para codificarlos, el código de mínima redundancia es aquel que emplea la cantidad media menor de dígitos para representar cada mensaje. En este sentido en realidad debiera hablarse de código óptimo. Para construir la distancia de Jensen – Shanon se trabaja de la siguiente forma [12].

Sea una variable aleatoria discreta X que toma N valores distintos sobre un conjunto $\Omega_N = \{\omega_1, \dots, \omega_N\}$. Se toma $\hat{X}$una muestra independiente e idénticamente distribuida (i.i.d). Esto quiere decir que cada valor es tomado como el de una variable aleatoria independiente de las otras e igualmente distribuida. Se supone que para seleccionar cada valor se toma cualquiera de dos distribuciones conocidas P y Q pero que no se sabe cuándo ha sido usada una y cuándo otra aunque la probabilidad de uso sea igual. Se quiere ahora nombrar la cantidad media menor de dígitos para representar los mensajes $X = \omega_i$ con $i = 1, \dots, N$. A esa cantidad promedio se la llama κ y claramente dependerá de una distribución R que dará la probabilidad r_i de cada mensaje $X = w_i$. Precisamente R será óptima cuando κ sea mínima y cada mínima media de dígitos para cada mensaje será $\kappa_i = -log r_i$ El valor esperado de κ si se da P es $E(\kappa, P)$ de forma que como $P\ y\ Q$ se dan al azar con equiprobabilidad queda $\kappa = \frac{1}{2}E(\kappa, P) + \frac{1}{2}E(\kappa, Q)$ y su mínimo valor ocurrirá cuando $R = \frac{1}{2}(P + Q)$ o sea cuando, en términos de la teoría de la información, $\kappa = H(R)$ entropía de R. Bajo tales circunstancias $H(R) = \frac{1}{2}H(P) + \frac{1}{2}H(Q)$ ya que $P\ y\ Q$ debieran ser iguales. Pero si no son exactamente iguales tendrá que ocurrir que son próximas y entonces $H(R) \approx \frac{1}{2}H(P) + \frac{1}{2}H(Q)$ con lo cual se puede definir la expresión

$$D_{PQ}^2 = 2H(R) - H(P) - H(Q)$$

Y entonces

$$\sqrt{D_{PQ}^2} = \left(2H(R) - H(P) - H(Q)\right)^{\frac{1}{2}}$$

cumple con las condiciones matemáticas para ser una distancia entre P y Q [12]. De acuerdo a ello queda:

$$D_{PQ} = \left(\sum_{i=1}^{N} p_i \log p_i + q_i \log q_i\right)^{\frac{1}{2}}$$

como expresión de la distancia entre dos distribuciones de probabilidad. En términos estadísticos frente a dos muestras de igual tamaño esto da una medida de que hayan sido tomadas bajo igual distribución de probabilidad. En los términos a los que se aplica el concepto cuanto más pequeño sea κ mayor es la proximidad probable entre ambos microbiomas. Esta distancia es la que usa el algoritmo PAM (Partitioning around medoids) utilizado para este caso directamente de la biblioteca cluster de R. El medoide es el elemento para el cual la disimilitud promedio con todos los objetos en el conglomerado es mínima. En realidad, el algoritmo PAM minimiza la suma de disimilitudes en vez de la disimilitud promedio. Se estudió también, con la idea de programarlo como alternativa más rápida si resultara necesaria, el algoritmo propuesto en [14] que trabaja con medoides pero en forma similar a k-means.

Para definir el número óptimo de clusters con los que trabajar en [6] se utiliza el índice de Calinski-Harabasz [15] que se construye de la siguiente forma:

$$CH = \frac{\frac{B_k}{k-1}}{\frac{W_k}{n-k}}$$

Aquí B_k es la suma de las distancias al cuadrado de todos los elementos *i* y *j* que no pertenecen al mismo cluster, W_k es la suma de los cuadrados de las distancias de todos los elementos *i* y *j* que pertenecen al mismo cluster, *n* es el número de elementos a clasificar y *k* la cantidad seleccionada de clusters. Utilizando el comando nclusters de la biblioteca clusterSim de R se puede programar el testeo de CH para distintos valores de *k* a fin de hallar la cantidad óptima de clusters.

Con el número de clusters obtenido de esta forma se entra entonces en el algoritmo PAM obteniendo los enterotipos en cuestión que resultan los distintos clusters. Desde el punto de vista computacional la consistencia de tal agrupamiento puede medirse con el índice Silhouette [16] Cada microbioma tiene un índice Silhouette $S(i)$ dónde i es el número de microbioma referido. Este índice varía entre -1 y 1: $-1 \leq S(i) \leq 1$. Además si $S(i) \approx 1$ el microbioma está muy bien clasificado dentro del cluster, si $S(i) \approx 0$ el microbioma estará tan bien en ese cluster como estaría en alguno de los otros y en la medida que $S(i) \approx -1$ el microbioma estará mal incorporado al cluster. Para cada cluster se toma el promedio de sus índices Silhouettes lo que permite comparar la bondad de los clusters entre sí. Además es posible obtener un índice Silhouette de toda la clusterización promediando los índices de los respectivos microbiomas . La bondad de una clasificación es mayor cuanto más cercano a 1 es el índice Silhouette general.

El análisis de componentes principales para reducir las dimensiones que caracterizan a cada microbioma es la última técnica utilizada. Tiene tanto objetivos de análisis clínico como de visualización gráfica. Las principales componentes se correlacionan fuertemente con aquellos microorganismos y grupos de microorganismos importantes en la determinación del enterotipo lo cual facilita la apreciación clínica. Por otra parte la visualización de los grupos obtenidos permite evaluar de modo práctico la diferenciación de cada cluster y su precisión respecto de la caracterización de una patología.

La segunda etapa del trabajo consistió en el procesamiento y análisis de un conjunto de microbiomas de pacientes relacionados con la Enfermedad de Crohn [9] ya referenciado más arriba. Este conjunto está integrado por dos lecturas de cada secuencia obtenida razón por la cual se contó con dos distribuciones de abundancia por paciente una vez realizada la asignación taxonómica. Esta tarea se desarrolló utilizando SUPERFOCUS. A efecto de hacer comparables las distribuciones de abundancias se tuvieron en cuenta 277 microorganismos distintos presentes en todas o en algunas de ellas solamente. En cada microbioma se despreciaron los microorganismos cuya concentración estuviera por debajo del porcentaje 0.01% Luego se usaron scripts realizados en R para leer los distintos archivos obtenidos y armar a partir de ellos la matriz de distancias en cada caso. Como se comentó más arriba la distancia utilizada fue la de Jensen-Shanon. Se aplicó luego el algoritmo PAM para hallar los clusters y se realizó un estudio del índice Silhouette obtenido según el número de clusters que se deseara establecer. Dado que las distribuciones de abundancia correspondientes a las lecturas 1 y 2 de la muestra se detallaron para el nivel de especies se consideró posible realizar un promedio de ambas en el caso de 10 pacientes pues el numerado con el 11 y clasificado previamente como sano poseía una sola lectura. También para estas distribuciones de abundancias promedio se realizó la determinación de clusters y el estudio del índice Silhouette. A continuación se evaluó la correspondencia de los agrupamientos obtenidos con la clasificación aportada por la apreciación clínica. Finalmente por medio del software

Infostat se calcularon las componentes principales a efecto de reducir las variables y obtener por un lado gráficas de los agrupamientos a la vez que para analizar la correlación de las mismas con la abundancia de los microorganismos.
Esta asociación evaluar la relación principal de ciertas especies con la presencia de la patología.

Resultados

En cuanto al proceso del primer conjunto de datos su resultado principal fue la determinación de una guía de trabajo y el entrenamiento en el uso y la sensibilidad de los algoritmos y programas aplicados.

En el caso del segundo conjunto se logró aplicar una metodología capaz de producir las distribuciones de abundancia por clasificación taxonómica y se evaluaron las dificultades presentadas para obtener tales distribuciones según la funcionalidad metabólica. A partir de la distribución de abundancia según taxonomía se desarrolló el proceso según la guía de trabajo establecida.

Los resultados obtenidos para los agrupamientos de la primera y la segunda lectura de las secuencias se muestran en las Figuras 1 y 2.

Figura 1

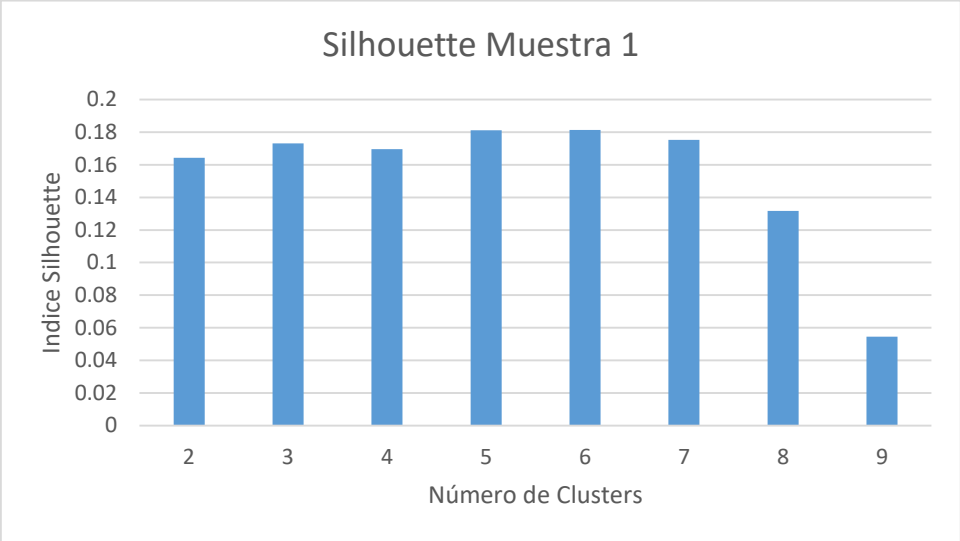
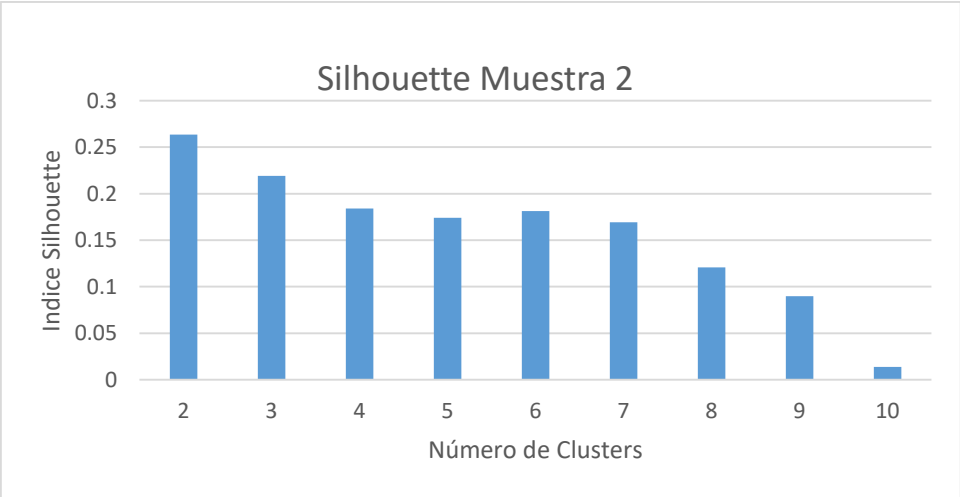
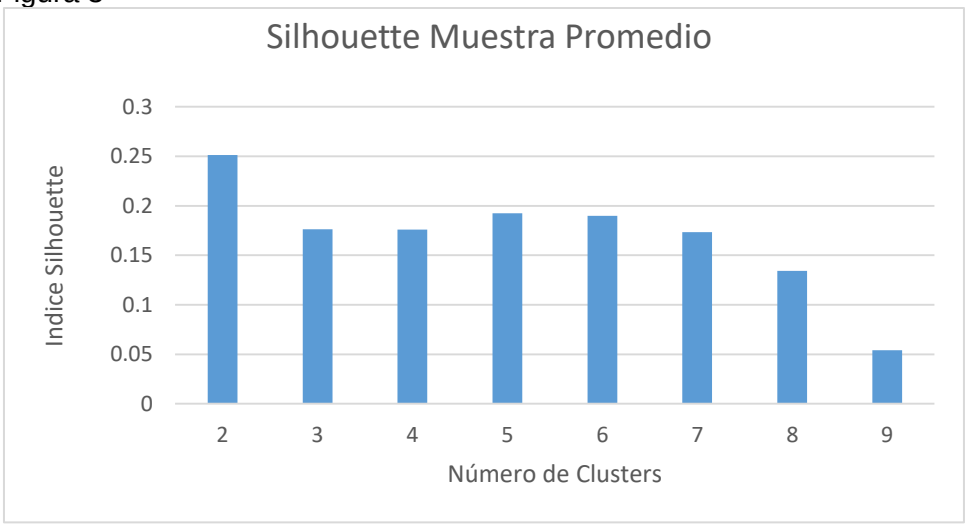


Figura 2



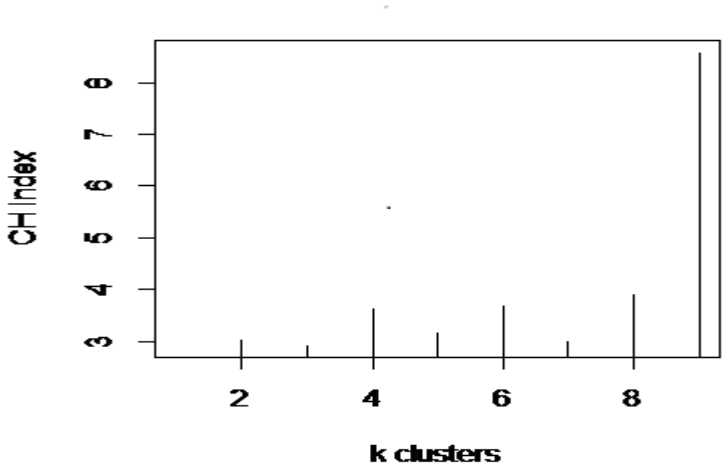
Se observa que para los microbiomas del Conjunto 1 establecidos a partir de las lecturas de las secuencias rotuladas con 1, no se registran índices Silhouette significativamente distintos entre la determinación de 2 hasta la de 7 clusters. No ocurre así en el caso correspondiente a las secuencias rotuladas con 2, asociadas al uso de distinto "primer" en la secuenciación, para el cual el agrupamiento en 2 clusters resulta el de mayor consistencia. Dado que se utiliza el nivel especie para las distribuciones de abundancia provenientes de ambas lecturas se consideró posible utilizar unas distribuciones construidas a partir de los promedios de abundancia. Con ellas se obtuvo el análisis de valores Silhouette que se muestra en la Figura 3.

Figura 3



Se observa aquí que el agrupamiento en 2 clusters o enterotipos resulta el más consistente. Sin embargo no deja de llamar la atención que el índice Silhouette óptimo sólo alcanza 0.25 lo que sugiere que la consistencia del agrupamiento así realizado es de todas formas débil. Es interesante observar que si se calcula el índice de Calinsky-Harabasz para determinar la cantidad de clusters óptima a determinar se obtiene la Figura 4

Figura 4

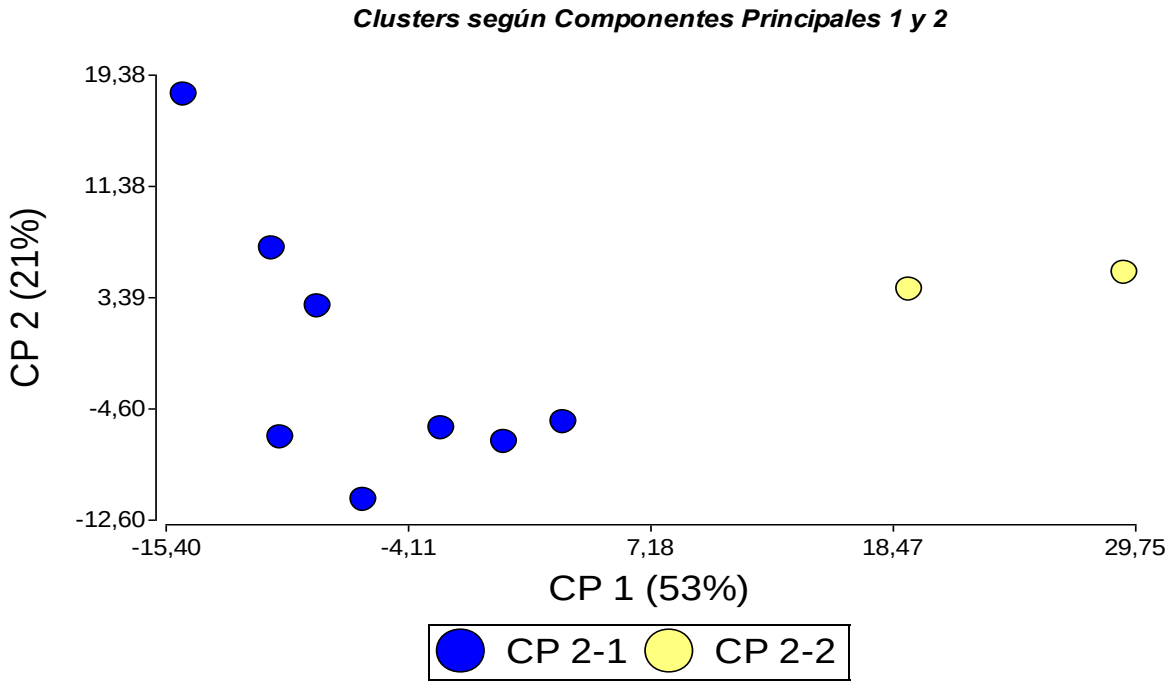


Estos resultados revelan que el agrupamiento en nueve clusters, sobre 10 microbiomas considerados, sería óptimo según el criterio de Calinsky-Harabasz. Esto puede deberse a la baja cantidad de microbiomas que integran la muestra pero refuerza la idea de la baja consistencia de la partición en clusters.

Con las reservas del caso se consideró entonces el agrupamiento en dos clusters como la mejor alternativa disponible considerándose que al menos estos podrían corresponderse con la clasificación en sano o enfermo que se conocía.

A continuación se realizó el análisis de componentes principales. Se obtuvo una primera componente CP1 que explica el 53% de la información y una segunda componente CP2 que condensa el 21% de la misma. La Figura 5 muestra los microbiomas coloreados según el cluster sobre el espacio bidimensional determinado por ambas componentes.

Figura 5



Se observan claramente los dos clusters bien diferenciados en colores azul y amarillo respectivamente.

El análisis acerca de la correlación de las dos primeras componentes principales con cada una de las variables (microorganismos) presentes en la distribución de abundancia arrojó una alta correlación en los casos expuestos en las Tablas 1 y 2

Tabla 1

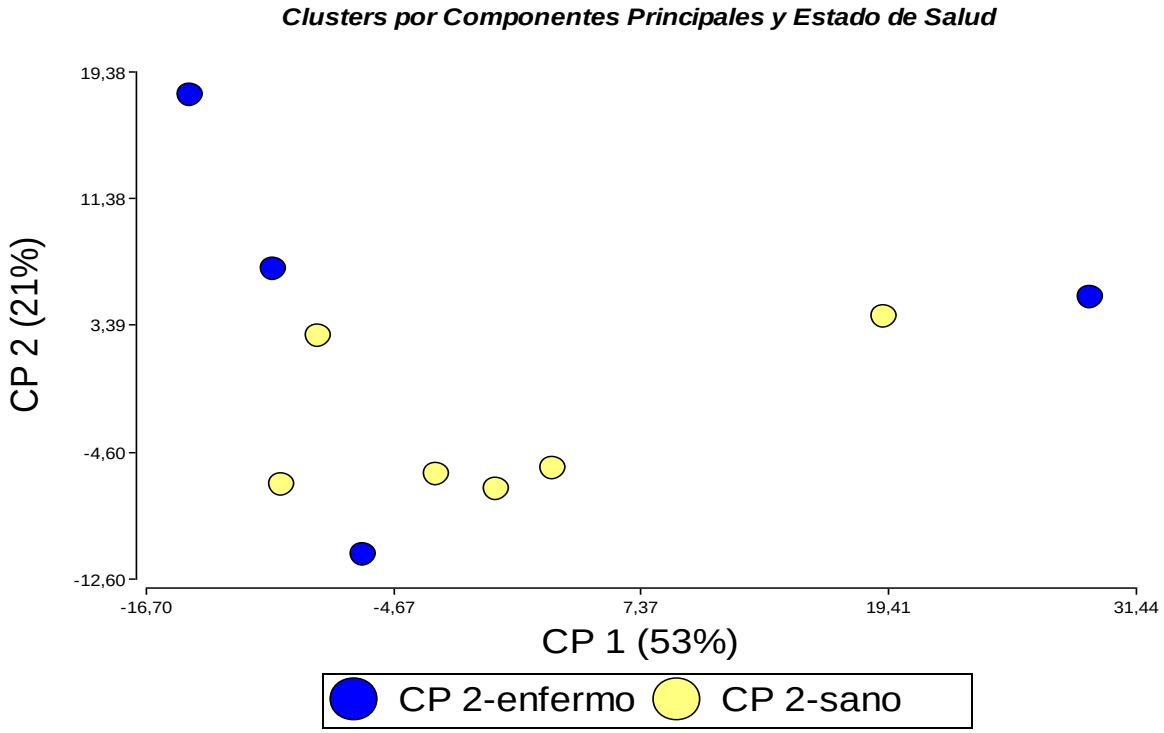
CP1	Alta Correlación
[Ruminococcus]_obeum	-0,92
Bacteroides_helcogenes	0,89
Bacteroides_salanitronis	0,83
Bacteroides_thetaiotaomicr..	0,92
Bacteroides_vulgatus	0,97
Bacteroides_xylanisolvens	0,96
Kitasatospora_setae	0,81
Porphyromonas_gingivalis	0,96
Prevotella_denticola	0,85

Tabla 2

CP2	Alta Correlación
butyrate-producing_bacteri..	0,75
Clostridium_saccharolyticu..	0,75
Roseburia_hominis	-0,75
Roseburia_intestinalis	-0,79

En particular se observa que la primera componente principal está altamente correlacionada con distintas Bacteroides y que la segunda se correlaciona con Clostridium. Esto es compatible con la observación de la bibliografía respecto de la Enfermedad de Crohn pues se ha encontrado que la manifestación de la enfermedad está asociada a una disminución de las especies bacterianas (Bacteroides y Clostridium) en la mucosa colónica. [17]. De acuerdo a ello sería esperable que los microbiomas de pacientes enfermos sean ubicados en el gráfico de acuerdo a bajos valores de las componentes principales 1 y 2 dada su correlación lineal directa con la abundancia de esas especies. Asimismo se observa que la componente CP1 mantiene una alta correlación lineal en este caso inversa con Ruminococcus que también pertenece a la clase Clostridia. Por tal motivo debiera esperarse que los microbiomas de pacientes enfermos se ubicaran en la gráfica adoptando valores altos del eje CP1. Estos fenómenos son los que se presentan en la Figura 6 en la cual se han coloreado distinto los pacientes sanos (amarillo) y los enfermos (azul)

Figura 6



Como puede observarse los agrupamientos realizados utilizando las distribuciones de abundancia elaboradas con criterio taxonómico no han logrado representar adecuadamente ya no una gradación que mostrara estadíos de enfermedad sino que tampoco han podido seguir la clasificación clínica enfermo-no enfermo obtenida por las vías diagnósticas usuales.

Conclusiones

En principio luego de los ensayos realizados con el primer conjunto, integrado directamente por distribuciones de abundancia ya obtenidas, se ha logrado establecer una secuencia o “pipeline” de tareas que ya se aplicó con el segundo conjunto relativo a la Enfermedad de Crohn.

También se ha conseguido procesar cada microbioma a partir de las secuencias originales en formato fastq adquiriendo conocimiento y operatividad sobre cambios de formato, programación para lectura y ordenamiento de conjuntos. El grupo en general ha profundizado sus conocimientos sobre el sistema operativo Linux y la programación del lenguaje R. Con respecto a la obtención de las distribuciones de abundancia taxonómica y funcional, se han presentado problemas pendientes de resolución con esta última, aunque en la primera fase del trabajo las pruebas pudieron realizarse adecuadamente. Luego se presentó un desajuste entre el sistema Linux y el programa SUPERFOCUS que deberá ser resuelto en la continuidad de la línea de investigación en un nuevo proyecto del programa de incentivos.

La conclusión más importante que puede extraerse es que el trabajo confirmó la idea de que los agrupamientos en base a taxonomía no parecen ser apropiados para establecer categorías o estadíos de enfermedad, al menos en el caso concreto aquí comprobado de la Enfermedad de Crohn y que deben profundizarse hacia el futuro los ensayos que utilicen distribuciones basadas en la funcionalidad metabólica de los genes secuenciados. Al respecto la bibliografía también sugiere que el estudio metagenómico

de biomarcadores funcionales arroja mejores resultados, hallándose correlación entre algunos de ellos y características del paciente. Sin perjuicio de esto ha sido reportado, en múltiples casos, contaminación del microbioma con el ADN del portador por lo que el desarrollo de mejores métodos experimentales es mandatorio para arribar a conclusiones [3].

Como reiteradamente se cita en el bibliografía [8], el microbioma adquiere distintas características y presenta sensibles variaciones por sexo, edad, origen étnico, alimentación al menos cuando se estudian sus aspectos taxonómicos. La variabilidad resulta menor cuando se considera el punto de vista funcional aunque esto tiene indudables características locales cuyo análisis solo podrá hacerse en cuenta si se toman adecuadas muestras en nuestro propio medio en vez de trabajar con muestras obtenidas de repositorios internacionales. Se abre entonces una etapa de ampliación de la tarea que deberá consistir en tomar muestras propias, enviarlas a secuenciar, trabajar con esa secuenciación para obtener los perfiles por funcionalidad y taxonomía, y establecer finalmente enterotipos para personas sanas y características de sus alteraciones en las personas enfermas. Tal etapa ya ha sido proyectada con la incorporación de médicos con formación específica en gastroenterología y cáncer para seguir ajustando la categorización computacional a la clínica tanto en el caso de la Enfermedad de Crohn como en el cáncer de colon.

Bibliografía

[1] Wang W-L, Xu S-Y, Ren Z-G, Tao L, Jiang J-W, Zheng S-S. Application of metagenomics in the human gut microbiome. World J Gastroenterol [Internet]. 2015 Jan 21 [cited 2015 Oct 13];21(3):803–14. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4299332&tool=pmcentrez&rendertype=abstract

[2] Weinstock GM. (2012) Genomic approaches to studyng the human microbiota. Nature Vol 489 250-256

[3] Shirley B, Alejandra V-L, Fernanda C-G, Karina R, Samuel C-Q, Xavier S, et al. Combining metagenomics, metatranscriptomics and viromics to explore novel microbial interactions: Towards a systems-level understanding of human microbiomes. Comput Struct Biotechnol J [Internet]. 2015 Jun [cited 2015 Jun 14];13:390–401. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4484546&tool=pmcentrez&rendertype=abstract

[4] Ngom-Bru, Catherine and Barretto, Caroline. Gut microbiota: methodological aspects to describe taxonomy and functionality. Briefings in Informatics. Vol3 NO 6. 747-750

[5] Statnikov, A. et al. A comprehensive evaluation of multicategory classification methods for microbiomic data. Microbiome 2013. 1:11

[6] https://www.r-project.org/

[7] www.infostat.com.ar

[8] Arumugam, M et al. Enterotypes of the human gut microbiome. Nature 2011 may 12; 473(7346): 174-180. doi:10.1038/nature09944

[9] Morgan, XC. et al. (2012) Dysfunction of the intestinal microbiome in inflammatory bowel disease and treatment. Genome Biology 2012, 13:R79

[10] https://edwards.sdsu.edu/SUPERFOCUS

[11] http://theseed.org

[12] Endres, D y Schindeling,J. A New Metric for Probability Distributions. IEEE. Transactions on Information Theory. Vol. 49 NO.7. 2003.

[13] Huffman, D. A Method for the Construction of Minimum-Redundancy Codes. Proceedings of the I.R.E. 1952. 1098.

[14] Hae-Sang Park, Chi-Hyuck Jun. A simple and fast algorithm for K-medoids clustering. Expert Systems with Applications 36 (2009) 3336-3341.
[15] Calinski, T., and J. Harabasz. "A dendrite method for cluster analysis." Communications in Statistics. Vol. 3, No. 1, 1974, pp. 1–27.
[16] Rousseeuw, P. Silhouettes: a graphical aid to the interpretation ans validation of cluster analysis. Journal of Computational and Applied Mathematics 20 (1987) 53-65.
[17] Ray K. IBD. Understanding gut microbiota in new-onset Crohn's disease. Nat Rev Gastroenterol Hepatol [Internet]. Nature Publishing Group; 2014;11(5):268. Available from: http://dx.doi.org/10.1038/nrgastro.2014.45\nhttp://www.nature.com/doifinder/10.1038/nrgastro.2014.45\nhttp://www.ncbi.nlm.nih.gov/pubmed/24662277
[18] Amiriam ES. Petrosino JF. Ajami NJ. Liu Y. Mims MP. Y Scheurer ME. (2013) Potential
role of gastrointestinal microbiota composition in prostate cancer risk. Infectious Agents and Cancer. 2013 8:42
[19]Burns MB, Lynch J, Starr TK, Knights D, Blekhman R. Virulence genes are a signature of the microbiome in the colorectal tumor microenvironment. Genome Med [Internet]. 2015 Jan [cited 2016 Feb 6];7(1):55. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4499914&tool=pmcentrez&rendertype=abstract
[20]Candela M, Turroni S, Biagi E, Carbonero F, Rampelli S, Fiorentini C, et al. Inflammation and colorectal cancer, when microbiota-host mutualism breaks. World J Gastroenterol [Internet]. 2014 Jan 28 [cited 2016 Jan 14];20(4):908–22. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3921544&tool=pmcentrez&rendertype=abstract
[21]Cho M, Carter J, Harari S, Pei Z. The interrelationships of the gut microbiome and inflammation in colorectal carcinogenesis. Clin Lab Med [Internet]. 2014 Dec [cited 2016 Feb 6];34(4):699–710. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4254421&tool=pmcentrez&rendertype=abstract
[22]Erickson AR, Cantarel BL, Lamendella R, Darzi Y, Mongodin EF, y cols. Integrated Metagenomics/Metaproteomics Reveals Human Host-Microbiota Signatures of Crohn's Disease. (2012) PLoS ONE 7(11): e49138. doi:10.1371/journal.pone.0049138
[23]Eckburg PB. Bik EM. Bernstein CN. Purdom E. Dethlefsen L. Sargent M. Gill SR. Nelson KE. Y Relman DA. (2005) Diversity of the Human Intestinal Microbial Flora. Science Vol 308 1635-1638
[24]Garcia TP. Müller S. Carroll RJ. y Walzem R. Identification of important regressor groups, subgroups and individuals via regularization methods: application to gut microbiome data (2013) Bioinformatics 2013 1-7
[25]Feng, Q et al. Gut microbiome development along the colorectal adenoma-carcinoma sequence. (2015) Nature Communications.6:6528 doi 10.1038/ncomms7528
[26]Fosso Bruno, Marzano M. y Santamaria, M e-DNA Meta-Barcoding: From NGS Raw Data to taxonomic Profiling. (2015) RNA Bioinformatics. Methods in Molecular Biology. Vol 1269 DOI 10.1007/978-1-4939-2291-8_16 Springer Science+Business
[27]Garcia, TP. et al. (2013) Identification of important regressor groups, subgroups and individuals via regularization methods: application to gut microbiome data. Bioinformatics Advance Access, 15 Nov 2013, pages 1-7, DOI:10.1093/bioinformatics/btt608
[28]Gill SR. Pop M. DeBoy RT. Eckburg PB. Turnbaugh PJ. Samuel BS. Gordon JI. Relman DA. Fraser-Liggett CM. y Nelson KE. (2006) Metagenomic Analysis of the Human Distal Gut Microbiome. Science Vol 312 1355- 1358
[29]Gradel KO, Nielsen HL, Schønheyder HC, Ejlertsen T, Kristensen B, Nielsen. (2009) Increased short- and long-term risk of inflammatory bowel disease after salmonella or

campylobacter gastroenteritis. Gastroenterology. 137(2):495.
[30]F. Guarner, J.R. Malagelada, (2003) Gut flora in health and disease. The Lancet, 361(9356):512-519. http://dx.doi.org/10.1016/S0140-6736(03)12489-0.
[31]Junhai et al. Diet, microbiota, and microbial metabolites in colon cancer risk in rural Africans and African Americans. Am. J. Clin. Nutr. 2013. 98. 11-120.
[32]Klindworth A. Pruesse E Schweer T. Peplies J. Quast C. Horn M. y Glöckner O. (2013) Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies. Nucleic Acids Research Vol. 41(1) :e1-e1
[33]Knights D. Costello E K. y Knight R. (2011) Supervised classification of human microbiota. FEMS Microbiol Rev 35 343-359
[34]Knights D, Ward TL, McKinlay CE, Miller H, Gonzalez A, McDonald D, et al. Rethinking "enterotypes". Cell Host Microbe [Internet]. 2014 Oct 8 [cited 2016 Feb 6];16(4):433–7. Available from: http://www.ncbi.nlm.nih.gov/pubmed/25299329
[35]Kuczynski J. Stombaugh J. Walters WA. González A. Caporaso G. y Knight R. (2011) Using QIIME to Analyze 16S rRNA Gene Sequences from Microbial Communities. Current Protocols in Bioinformatics 10.7.1-10.7.20,
[36]Lan Y. Kriete A. y Rosen GL (2013) Selecting age-related functional characteristics in
the human gut microbiome. Microbiome 2013 1:2
[37]Li K. Bihan M. Yooseph S. y Methe BA. (2012) Analyses of the Microbial Diversity across the Human Microbiome PLoS ONE 7(6): e32118.
[38]Mackenzie, B W, Waite, D W, Taylor, M W. Evaluating variation in human gut microbiota profiles due to DNA extraction method and inter-subject differences. (2015) Frontiers in Microbiology. Vol 6. Article 130.
[39]McMurdie PJ. y Holmes S. (2013) phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data. PLoS ONE 8(4): e61217.
[40]Molodecky N.A., Soon I.S., Rabi D.M., y cols. Increasing incidence and prevalence of the inflammatory bowel diseases with time, based on systematic review. (2012) Gastroenterology 2012(142):46-54.
[41]Morgan XC. Segata N. y Curtis Huttenhower. (2013) Biodiversity and functional genomics in the human microbiome. Trends in Genetics. Vol 29 No. 1
[42]Parks, D H y Beiko, R G. Identifying biologically relevant differences between metagenomic communities. (2010) Genome Analysis. Oxford University Press.
[43]Perez-Brocal V et al. Study of the Viral and Microbial Communities Associated With Crohn Disease: A Metagenomic Approach. (2013) Clinical and Transactional Gastroenterology. 4, e36; doi : 10,1038/ctg.2013.9
[44]Robertson CE. Harris JK. Wagner BD. Granger D. Browne K. Tatem B. , Feazel LM.
Park K. Pace NR. Y Frank DN. (2013) Explicet: graphical user interface software for metadata-driven management, analysis and visualization of microbiome data. Bioinformatics Vol 29 No 23 3100-3101.
[45]Sohn, MB. An, Lingling. Pookhao, Naruekamol. Li, Qike. (2014) Accurate genome relative abundance estimation for closely related species in a metagenomic simple. BMC Bioinformatics 2014, 15:242
[46]Thorkildsen LT, Nwosu FC, Avershina E, Ricanek P, Perminow G, Brackmann S, Vatn MH, Rudi K. Dominant fecal microbiota in newly diagnosed untreated inflammatory bowel disease patients. (2013) Gastroenterol Research and Practice. 2013:636785. doi: 10.1155/2013/636785.

[47]Valles-colomer M, Darzi Y, Vieira-silva S, Falony G, Raes J, Unit M, et al. Meta-omics in IBD research: applications, challenges and guidelines. J Crohn's Colitis Adv Access. 2016;1–34.
[48]White, James R., Nagarajan, Niranjan y Pop, Mihai.(2009) Statistical Methods for detecting Differentialy Abundant Features in Clinical Metagenomic Samples. Plos Computational Biology Vol. 4 Issue 4 e1000352
[49]Weir TL, Manter DK, Sheflin AM, Barnett BA, Heuberger AL y cols. Stool Microbiome and Metabolome Differences between Colorectal Cancer Patients and Healthy Adults. (2013) PLoS ONE 8(8): e70803. doi:10.1371/journal.pone.0070803
[50]Wu, X et al. (2013) Comparative analysis of microbiome measurement platforms using latent variable structural equation modeling. BMC Bioinformatics 14:79
[51]Xin F. Chen J. Fung WK. y Li H. (2013) A logistic normal Multinomial Regression Model
for Microbiome Compositional Data Analysis. (2013) Biometrics doi: 10.1111/biom.12079
[52] Yang T, Owen JL, Lightfoot YL, Kladde MP, Mohamadzadeh M. Microbiota impact on the epigenetic regulation of colorectal cancer. Trends Mol Med [Internet]. 2013 Dec [cited 2016 Feb 2];19(12):714–25. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3851951&tool=pmcentrez&rendertype=abstract
[53]Zackular, JP. Baxter NT. Iverson KD. Sadler WD. Petrosino JF. Chen GY. Y Schloss PD. (2013) The Gut Microbiome Modulates Colon Tumorgenesis. mBio 4(6):e00692-13
[54]Zackular JP, Rogers MAM, Ruffin MT, Schloss PD. The human gut microbiome as a screening tool for colorectal cancer. Cancer Prev Res (Phila) [Internet]. 2014 Nov [cited 2015 Nov 25];7(11):1112–21. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4221363&tool=pmcentrez&rendertype=abstract

3. Cuerpo de Anexos

Anexo I
Anexo II
Anexo III
1-WICC2016_1
2-WICC2016_2
3-Raices2015
4-WICC2015_1
5-WICC2015_2
6-WICC2014
7-Unlam2016
Anexo IV
1-Minería de Datos para Análisis del Microbioma Humano.pdf WICC 2017
2- Póster WICC2016.pptx
3- Clasificación por Enterotipos y Grupos Ortólogos del Microbioma Humano con Métodos No Supervisados.docx WICC 2016
4- Técnicas de Minería de Datos Aplicadas al Procesamiento de ADN de Comunidades Microbiológicas.pdf WICC 2015
5- Investigación Ingeniería- Avances.pdf Revista Avances. UNLAM
6- Clustering y Ensambles de Árboles de Decisión Aplicados sobre el Microbioma Huamano.pdf
Anexo V