Trabajo de Fin de Grado

QiimeApp – Una plataforma web para análisis metagenómicos

QiimeApp – A web platform for metagenomic analysis

Víctor Juidías Rodríguez

La Laguna, 6 de junio de 2016

D. **Marcos Colebrook Santamaría**, con N.I.F. 43.787.808-V, profesor Titular de Universidad adscrito al Departamento de Ingeniería Informática y de Sistemas de la Universidad de La Laguna, como tutor

D. **José L. Roda García**, con N.I.F. 43.356.123-L, profesor Titular de Universidad adscrito al Departamento de Ingeniería Informática y de Sistemas de la Universidad de La Laguna, como cotutor

C E R T I F I C A N

Que la presente memoria titulada:

QiimeApp – Una plataforma web para análisis metagenómicos

ha sido realizada bajo su dirección por D. **Víctor Juidías Rodríguez**, con N.I.F. 54.057.108-R

Y para que así conste, en cumplimiento de la legislación vigente y a los efectos oportunos firman la presente en La Laguna a 6 de junio de 2016.

Agradecimientos

El presente trabajo, no habría sido posible sin la inestimable ayuda de numerosas personas, que con su colaboración y apoyo me han permitido desarrollar el proyecto hoy se presenta.

Me gustaría agradecer a los directores del proyecto el Dr. Marcos Colebrook y Dr. José Luis Roda, ambos profesores pertenecientes a la Escuela Superior de Ingeniería y Tecnología, Sección de Ingeniería Informática (ESIT–SII), y muy especialmente agradecer al Dr. Carlos Flores, investigador del Hospital Universitario Nuestra Señora de la Candelaria y del Grupo de Genómica Aplicada de La Universidad de La Laguna, por su comprensión y paciencia conmigo y por haberme podido guiar durante todo el desarrollo de este proyecto, solventando las dudas que iban surgiendo, para poder hacerlo posible.

También agradecerles la enorme ilusión transmitida acerca de este nuevo mundo, el de la Bioinformática, un mundo que ha resultado ser tan apasionante y con grandes perspectivas de futuro, de cara a la mejora de la calidad de vida de las personas, que no podemos dejar escapar la oportunidad de poder trabajar en él.

Licencia

Resumen

El objetivo de este trabajo ha sido desarrollar una aplicación bioinformática que sirva de ayuda para las investigaciones que son llevadas a cabo por el personal del Hospital Universitario Nuestra Señora de la Candelaria y la Unidad de Genómica Aplicada de la Universidad de La Laguna. Ante la falta de una adecuada herramienta encargada de procesar y llevar a cabo análisis metagenómicos sobre datos obtenidos de la secuenciación masiva de ADN, se presenta *QiimeApp – una plataforma web para análisis metagenómicos* como solución a esta problemática. Una aplicación web capaz de llevar a cabo este proceso de una manera automatizada, cómoda y transparente para el usuario.

Palabras clave: *ADN*, *Metagenómica, Gen 16S rRNA*

Abstract

The aim of this project has been to develop a bioinformatics application that it will help in the research experiments that are being carried out by staff of the *Hospital Universitario Nuestra Señora de la Candelaria* and Applied Genomics Unit at the University of La Laguna. Because of the lack of a proper tool responsible for processing and perform metagenomic analysis of data obtained from massive DNA sequencing, *QiimeApp – a web platform for metagenomics analysis* is presented as a solution to this problem. A web application capable of carrying out this process in an automated, simple and transparent way for the user.

Keywords: *DNA, Metagenomics, 16S rRNA Gene*

Índice General

Capítulo 1. Introducción **6**

1.1 La Metagenómica 6

1.2 La Bioinformática 11

Capítulo 2. Estado del Arte **13**

2.1 Introducción 13

2.2 Soluciones comerciales y no comerciales 13

Capítulo 3. Diseño de la aplicación **17**

3.1 Introducción 17

3.2 Diseño de la aplicación 20

Capítulo 4. Desarrollo de la aplicación **23**

4.1 Desarrollo del *pipeline* 23

4.1.1 Descripción de los scripts 25

4.1.2 Herramientas adicionales: PICRUST y FastQC 31

4.2 Desarrollo de la aplicación Web 32

Capítulo 5. Resultados **34**

5.1 Introducción 34

5.2 Pantalla de resultados 35

5.2.1 Resultados de QIIME 37

5.2.2 Resultados de PICRUST 41

5.2.3 Resultados de FastQC 43

Capítulo 6. Conclusiones y líneas futuras **44**

Capítulo 7. Presupuesto **47**

Capítulo 8. Summary and Conclusions **48**

Bibliografía **50**

Índice de figuras

Figura 1.1. Diferencia entre Genómica y Metagenómica 8
Figura 1.2. Diferentes procedimientos dentro de la Metagenómica 8
Figura 1.3. Proceso de análisis metagenómico 11
Figura 3.1. Índice con algunos de los scripts incluidos en QUIIME 17
Figura 3.2. Ejemplo de la ayuda de un script de QUIIME 18
Figura 3.3. Ejemplo de ejecución de un script de QUIIME 19
Figura 3.4. Esquema del diseño de la aplicación 22
Figura 4.1. Esquema del diseño del pipeline 25
Figura 4.2. Proceso de demultiplexación 27
Figura 4.3. Proceso de clustering 29
Figura 4.4. Esquema de diseño del nuevo pipeline 32
Figura 5.1. Pantalla de resultados 35
Figura 5.2. Pantalla de histogramas 36
Figura 5.3. Datos de resumen de QUIIME 36
Figura 5.4. Número de individuos en la muestra 37
Figura 5.5. Porcentaje de cada individuo de la muestra 38
Figura 5.6. Clasificación de los individuos según los sexos de la muestra 38
Figura 5.7. Taxonomía de cada individuo según los sexos de la muestra 39
Figura 5.8. Número de individuos clasificados por las razas de la muestra 39
Figura 5.9. Taxonomía de los individuos según las razas de la muestra 40
Figura 5.10. Visualización tridimensional de los resultados 40
Figura 5.11. Resultados de PICRUSt en diagrama de barras 41
Figura 5.12. Resultados de PICRUSt en formato de áreas 42
Figura 5.13. Taxonomía de las funciones según cada uno de los individuos 42
Figura 5.14. Representación de las muestras antes del filtrado 43
Figura 5.15. Representación de las muestras después del filtrado 43

Índice de tablas

Tabla 7.1. Presupuesto de la aplicación 47

Capítulo 1. Introducción

1.1 La Metagenómica

Gracias al desarrollo de los primeros microscopios en torno al siglo XVII, se pudieron llevar a cabo los primeros descubrimientos y estudios sobre microorganismos, de los que la humanidad se ha beneficiado enormemente desde entonces. A partir de este momento, el ser humano fue capaz de entender las razones por las cuales se producían procesos, que hasta entonces le eran totalmente ajenos y que eran llevados a cabo por estos pequeños seres vivos, incapaces de ser vistos por el ojo humano. Procesos tales como, porqué las uvas se podían convertir en vino o porqué los alimentos se estropeaban, entre otros tantos. Pero lo que fue realmente importante, fue saber que estos seres son los responsables de causar mortales enfermedades e infecciones. Por ello, la humanidad se ha visto obligada a dedicar grandes esfuerzos a estudiar estos pequeños organismos, de los que conocemos una muy pequeña parte de las millones de especies que existen en el mundo.

Todas ellas tienen en común la de codificar la información básica de su existencia en una macromolécula básica, el ácido desoxirribonucleico o **ADN**. De una manera simple, podemos decir que el ADN está compuesto a su vez por cuatro moléculas orgánicas básicas o unidades, los **nucleótidos** (comúnmente denotados como A, C, G y T). Por ello, conociendo su orden en la secuencia de ADN podemos diferenciar las distintas especies. Sustituciones de una base por otra en el material genético, pueden producir enfermedades o constituir la semilla para la posterior aparición de nuevas especies. Tenemos que tener en cuenta que la consecución de estos nucleótidos es el nivel más básico al que podemos describir los genes, las unidades básicas de funcionamiento de los seres vivos.

Los métodos tradicionales para el estudio de microorganismos necesitan primero de la obtención de las muestras de un ambiente en concreto, para

luego ser cultivadas y poder así, obtener la información concreta sobre los organismos que componen la muestra en estudio. Considerando que no todos los microorganismos son aptos para ser cultivados en laboratorio, se estima que, en torno al 99% de los posibles organismos presentes en la muestra no serán caracterizables mediante estos métodos tradicionales. Si, como media, tan sólo el 1% de los organismos de una muestra será cultivable, es fácil anticipar que los métodos tradicionales no han sido de gran utilidad en la caracterización de microorganismos. A pesar de ello, su uso en el ambiente hospitalario es rutinario para la detección de infecciones.

Hoy en día la **Metagenómica** [1], que es una rama de la Genómica, la ciencia encargada del estudio integro de la información genética de un individuo, lo que conocemos como genoma, es un campo que se ha ido abriendo paso poco a poco durante los últimos años, y que está suponiendo toda una revolución en las investigaciones científicas relacionadas con el estudio del material genético de los seres vivos.

La Metagenómica carece precisamente del problema anteriormente mencionado, al ser capaz de obtener la información del 100% de los organismos presentes en la muestra en estudio (ver Figura 1.1). La Metagenómica permite el estudio del material genético recogido directamente a partir de muestras de un ambiente concreto, sin la necesidad de tener que ser aisladas y cultivadas en un laboratorio. Es un campo emergente, que nos ha abierto las puertas al estudio de organismos que hasta ahora eran completamente desconocidos, e incluso, ser capaces de poder estudiar comunidades microbianas completas.

La Metagenómica es una de las nuevas tecnologías, que han surgido gracias al avance durante los últimos años, de los llamados **secuenciadores de nueva generación** (NGS, *Next Generation Sequencer*) [2], que permiten extraer y secuenciar, a bajo coste, el ADN recogido a partir de una muestra de cualquier ambiente.

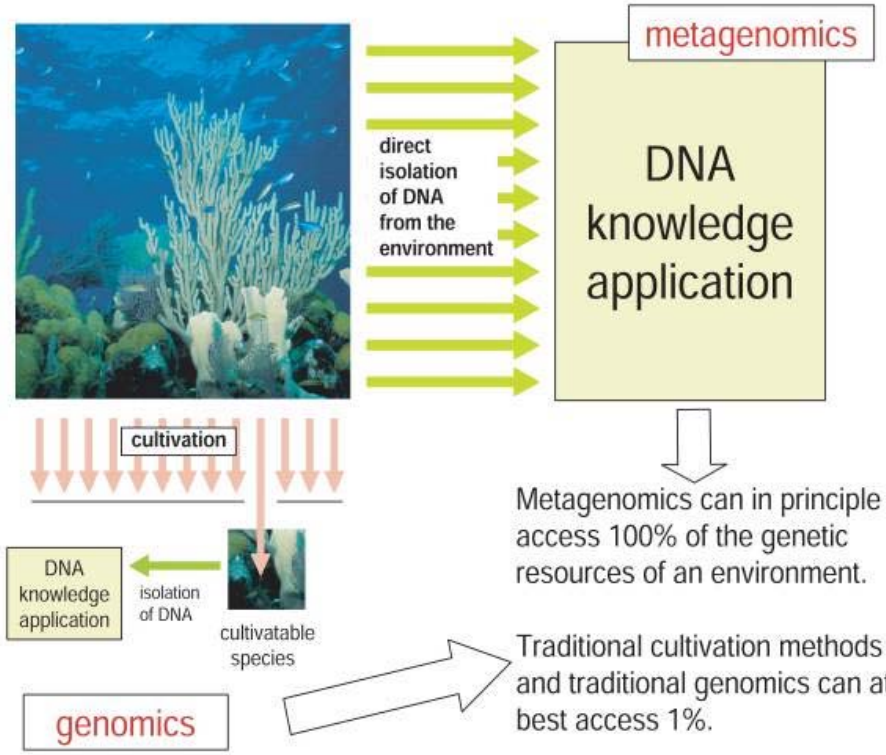


Figura 1.1. Diferencia entre Genómica y Metagenómica

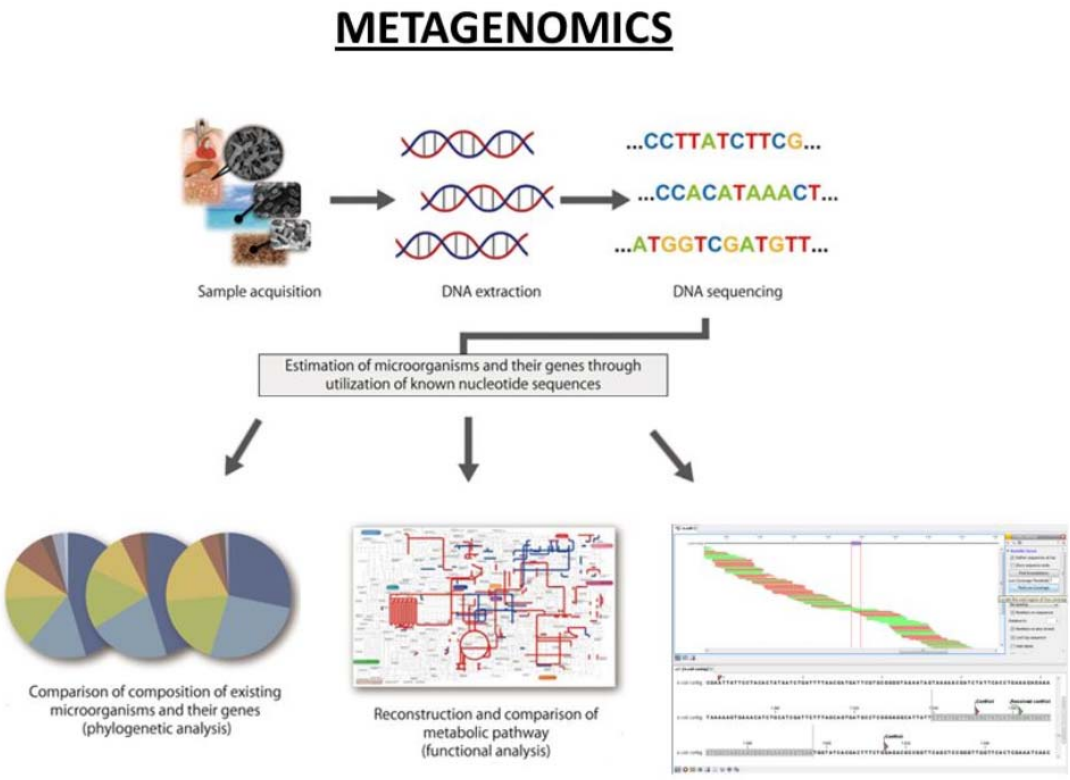


Figura 1.2. Diferentes procedimientos dentro de la Metagenómica

Secuenciadores como el PGM (Personal Genome Machine) [3] de propiedad de Thermo Fisher , o el potente MiSeq [4] de Illumina, , han permitido llevar a cabo grandes avances y que, con relativa facilidad cualquier estudio de investigación pueda ser capaz de generar cientos de miles de datos de una gran calidad en unas pocas horas

Podemos simplificar indicando que la metagenómica es una tecnología basada en la secuenciación y análisis de ADN que ha sido extraído directamente de diversos ambientes, como agua, suelo, saliva, tracto digestivo, etc. Además, la metagenómica, nos aporta la información acerca de la estructura de la comunidad microbiana, datos como qué especies la habitan, su diversidad o distribución. Por otro lado, también nos permite conocer las funciones desempeñadas dentro de dicha comunidad por los organismos que la componen (ver Figura 1.2).

En metagenómica, existen principalmente dos técnicas a la hora de llevar a cabo las investigaciones:

A) Por un lado, tenemos la **metagenómica dirigida**, que estudia los organismos presentes en una comunidad microbiana, su composición y su diversidad, pero llevando a cabo el análisis centrándose únicamente en uno varios marcadores genéticos muy específicos. Es decir, de todo el material genético extraído de la muestra, filtrará solamente aquellas secuencias que se correspondan con el gen o genes que se busca.

   El enfoque más utilizado a la hora de decantarse por usar metagenómica dirigida, es en base al gen que codifica la subunidad **16S del ARN ribosomal o simplemente 16S rRNA** [5]. Es un gen que está presente en la totalidad de los procariontes, es decir, aquellos organismos que están constituidos por células procariotas (incluyendo *bacteria*), y en el ADN mitocondrial de las células eucariotas (nótese que a las mitocondrias se les asume un origen evolutivo en un organismo ancestral relacionado con las bacterias).

   Sobre este tipo de enfoque, la metagenómica dirigida al gen 16S rRNA, más accesible por su bajo coste, es sobre el que se ha basado el presente proyecto para desarrollar la aplicación.

B) Por otro lado, la **metagenómica no dirigida o *shotgun***, no se centra en estudiar un gen en particular. Al contrario, lo que persigue es conocer todo el material genético de todos los organismos de la comunidad microbiana. Es decir, mientras que en la metagenómica dirigida nos centramos en una parte muy concreta de la muestra, para quedarnos sólo con aquellos organismos que contienen el gen que buscamos y nos olvidamos del resto, en la metagenómica *shotgun* lo que se busca es obtener una vista completa de todos los organismos presentes en la muestra y de sus genes. Por lo tanto, en esta técnica se secuencian directamente genomas microbianos completos. (ver Figura 1.3).

Las muestras en estudio contienen infinidad de microorganismos en distintas proporciones. Por ello, las especies más abundantes tendrán una representación mayor en la información que obtenemos al secuenciar el material genético. Para poder conseguir información suficiente sobre los organismos con una menor representación, se requiere disponer de gran cantidad de material de partida que asegure una cobertura suficiente. Utilizando metagenómica, nos aseguramos que estos organismos estarán representados, al menos, por pequeños fragmentos de su secuencia que, de lo contrario, pasarían completamente desapercibidos utilizando los métodos tradicionales de aislamiento y cultivo.

Como resultado de los análisis metagenómicos llevados a cabo, obtendremos miles, cuando no millones de datos, procedentes de la secuenciación del material genético de los organismos de la comunidad microbiana bajo estudio. El paso siguiente es poder ser capaces de procesar tal volumen de datos, para darles sentido y que la información que se pueda obtener de ellos sea realmente útil e interpretable en el contexto de la investigación en curso.

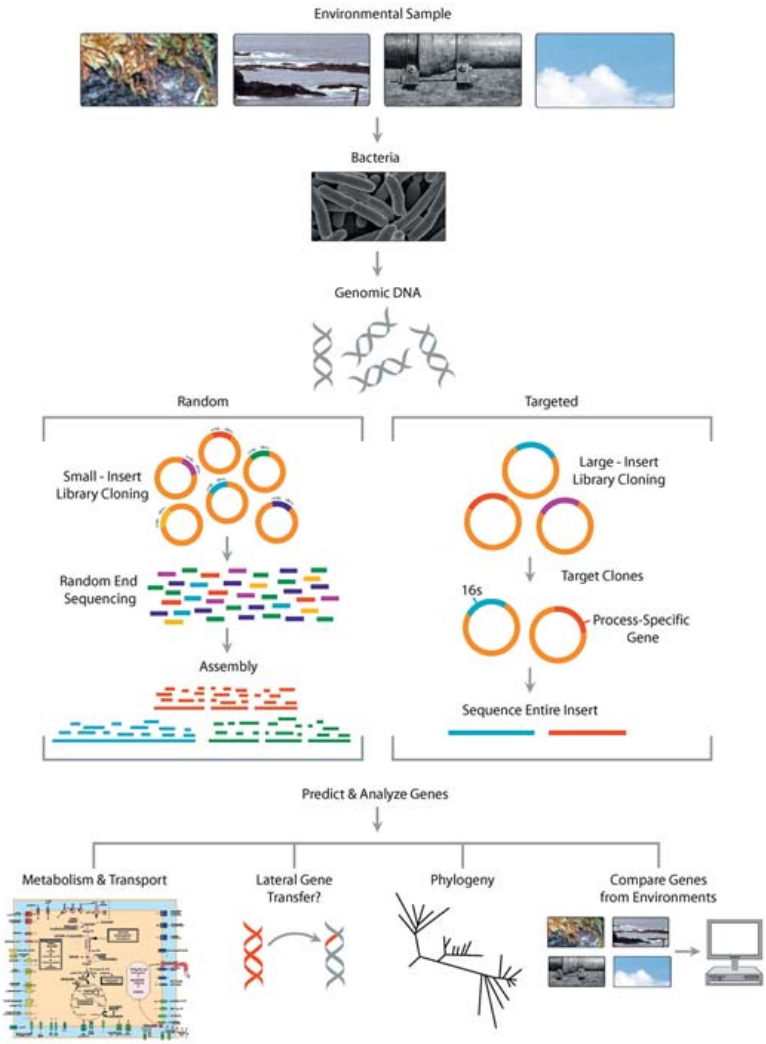


Figura 1.3. Proceso de análisis metagenómico

1.2 La Bioinformática

En este punto clave la **Bioinformática**, es decir, el conjunto de técnicas y herramientas informáticas desarrolladas con el fin de poder gestionar, analizar e interpretar datos biológicos.

Dentro los objetivos principales que tiene la Bioinformática encontramos, la **secuenciación del ADN** propiamente y **el alineamiento de secuencias**, que consiste en comparar dos o más secuencias de ADN, de ARN o de proteínas, para identificar regiones de similitud y poder establecer relaciones entre ellas.

Para lograr estos objetivos, en Bioinformática se utilizan técnicas tan diversas cómo, reconocimiento de patrones, minería de datos, algoritmos de *machine learning* y sobre todo técnicas de visualización de datos.

Como ejemplo de una herramienta bioinformática, en el contexto del grupo se desarrolló en el pasado la herramienta web **IonGAP** [6], un conjunto integrado de procesos, o pipeline, concebido para facilitar la secuenciación, análisis e interpretación de datos de genomas bacterianos y que está específicamente desarrollada para funcionar a partir de lecturas de secuenciadores de tipo Ion Torrent.

Capítulo 2.
Estado del Arte

2.1 Introducción

En el desarrollo de cualquier proyecto de investigación en el campo de la metagenómica, se depende en gran medida de la existencia de herramientas informáticas capaces de analizar grandes volúmenes de datos procedentes de la extracción del material genético de una muestra ambiental.

Afortunadamente, a la par que han ido apareciendo en el mercado la nueva generación de secuenciadores, se ha ido desarrollando en paralelo, un buen número de soluciones de software capaces de dar respuesta a la demanda de datos que necesitan ser analizados.

En el caso que nos concierne en este proyecto, son herramientas para Metagenómica dirigida al gen **16S rRNA**. Para el análisis de los miles de datos arrojados por los secuenciadores, existen varias soluciones disponibles en el mercado, desde herramientas con licencia comercial, pasando por las *open-source*, que son las más comunes entre la comunidad investigadora. Además, también existen algunas soluciones en la nube.

2.2 Soluciones comerciales y no comerciales

Empezando por las soluciones comerciales, existen varias, pero cada una ligada a la plataforma a la cuál pertenezca el secuenciador. Tanto Illumina como Thermo Fisher proveen de software especializado en análisis metagenómico, el cuál funciona adecuadamente sólo cuando se trabaja conjuntamente con sus propios secuenciadores. Por contrapartida, su gran desventaja es su poca flexibilidad. Son herramientas diseñadas para trabajar por y para estas plataformas, por ello es casi imposible intentar combinarlas con otras herramientas para trabajar de manera conjunta y obtener una información más completa. Otra desventaja, como es obvio, es que para poder utilizarlas se han de adquirir las licencias de uso. Además, existen otras

herramientas con licencia *open-source*, con las se pueden obtener los mismos resultados, lo que lleva a pensar como es lógico, que se termine optando por éstas últimas.

En cuanto a las soluciones *online* o basadas en la nube, destaca como principal ventaja, no tener que instalar ningún tipo de herramienta ni dependencias en los terminales de trabajo y no tener que lidiar con problemas de instalaciones o de funcionamiento, o de disponer de un mínimo de requerimientos en cuanto al hardware de la máquina se refiere, requisito indispensable a la hora de poder ejecutar análisis sobre grandes cantidades de datos. Nótese que el resultado de secuenciar una muestra típica conlleva generar ficheros de datos que oscilan en la escala de las varias decenas de GB para proyectos pequeños, a la escala de la decena de TB en el contexto de proyectos con un tamaño medio o grande. Por ello estas plataformas online, ya tienen incorporado todo el hardware necesario para trabajar con este volumen de datos.

Para este apartado vamos a diferenciar entre las herramientas online disponibles, las cuales ya vienen con software preinstalado, y nosotros sólo tenemos que subir los ficheros de datos y seleccionar los parámetros del análisis. Y por otro lado las soluciones en la nube basadas en lo que hoy día se denomina como *Infrastructure as a Service* (IaaS), soluciones como la nube de Amazon o la de Google.

Como ejemplos de plataformas online para llevar a cabo análisis metagenómicos dirigidos al gen 16S, podemos encontrar, por ejemplo, el **EBI metagenomics webserver** [7], una herramienta gestionada por *European Molecular Biology Laboratory-European Bioinformatics Institute* (EMBL-EBI), **SILVA NGS** [8] **y MG-RAST** [9].

Pero este tipo de plataformas cuentan con una enorme desventaja y es que en su contra juega, la desconfianza generalizada que existe dentro del mundo de la investigación. Es decir, los investigadores o usuarios potenciales de la aplicación, tienen invertido mucho dinero, tiempo y esfuerzo para, por un lado, poder conseguir la adjudicación del proyecto deseado, conseguir la financiación requerida para el proyecto, y por último, conseguir todas los datos necesarios, con una calidad apropiada que les permita obtener los resultados deseados al concluir el estudio. En este último paso, entran en

juego los secuenciadores de última generación que han revolucionado la forma en la que hasta ahora se trabajaba en Genómica, reduciendo notablemente el coste en recursos y tiempo que se necesitaba para poder manipular las muestras en el laboratorio. Por tanto, se entiende que, desde su punto de vista, exista cierta desconfianza al dejar los datos, que tanto les han costado conseguir, a una plataforma de terceros, es decir que no es propia, y no tener la garantía de que no se va a hacer un uso indebido de los datos de su proyecto.

Además, este tipo de soluciones tiene otro punto en contra, y es el rendimiento. Estas plataformas funcionan con colas de prioridad, es decir, que no comenzarán a analizar los datos que acabas de subir, sino que los pondrán en una cola, y sólo cuando sea su turno, será cuando sean procesados. Cabe destacar también que los recursos de los que disponen este tipo servicios, están compartidos, con lo que nunca se va a poder disponer del 100% de los recursos disponibles para llevar a cabo los análisis pertinentes.

Esto sumado a lo anterior, hace que, al recurrir a este tipo de soluciones, el análisis de nuestros datos se prologue demasiado en el tiempo, como para considerar si quiera estas opciones viables, cuando andamos escasos precisamente de eso, de tiempo.

Por otro lado, tenemos también otro tipo de soluciones en la nube, que es diseñarlo nosotros mismos, en cualquiera de los servicios que hay disponibles hoy día en el mercado. Las opciones más comunes suelen ser la nube de Google o la de Amazon. Éstas nos permiten, que elijamos la configuración de hardware que mejor se adapte a nuestras necesidades de cada momento dependiendo del volumen de datos de nuestro proyecto que queremos analizar. Pero el problema recae en el precio de alquilar dichos servicios. La mayoría de las veces no sale rentable plantearse como una opción viable este tipo de servicios, que recordemos, no nos quita de tener que desarrollar nuestros propios *workflow* de trabajo.

Además, en cualquiera de las dos opciones anteriores, dependemos mucho de la velocidad de nuestro servicio de internet, puesto que debemos de transferir ficheros de decenas, cuando no de cientos de gigas, a través de la red hacia estas plataformas. Si carecemos de la velocidad adecuada para llevar

esto a cabo en un tiempo razonable, resulta inviable si quiera plantearnos cuál de las anteriores opciones es la mejor para nuestro proyecto.

Es por ello, que la opción finalmente escogida por la mayoría, sea la de desarrollar aplicaciones propias o recurrir a herramientas de carácter *open-source,* que pueden adaptar según sus necesidades, y así tener la garantía, de que nadie más que ellos son los dueños y señores de sus datos y sus proyectos.

En el apartado de herramientas *open-source* instalables en los terminales de trabajo, encontramos principalmente dos, **QIIME** [**10**] y **MOTHUR** [**11**]. Ambas herramientas *open-source*, que persiguen el objetivo de facilitar la labor de investigación, capaces de realizar análisis metagenómicos sobre las dos plataformas principales, Illumina e IonTorrent. MOTHUR ha sido desarrollada usando C/C++, mientras que QIIME es una herramienta implementada completamente en Python.

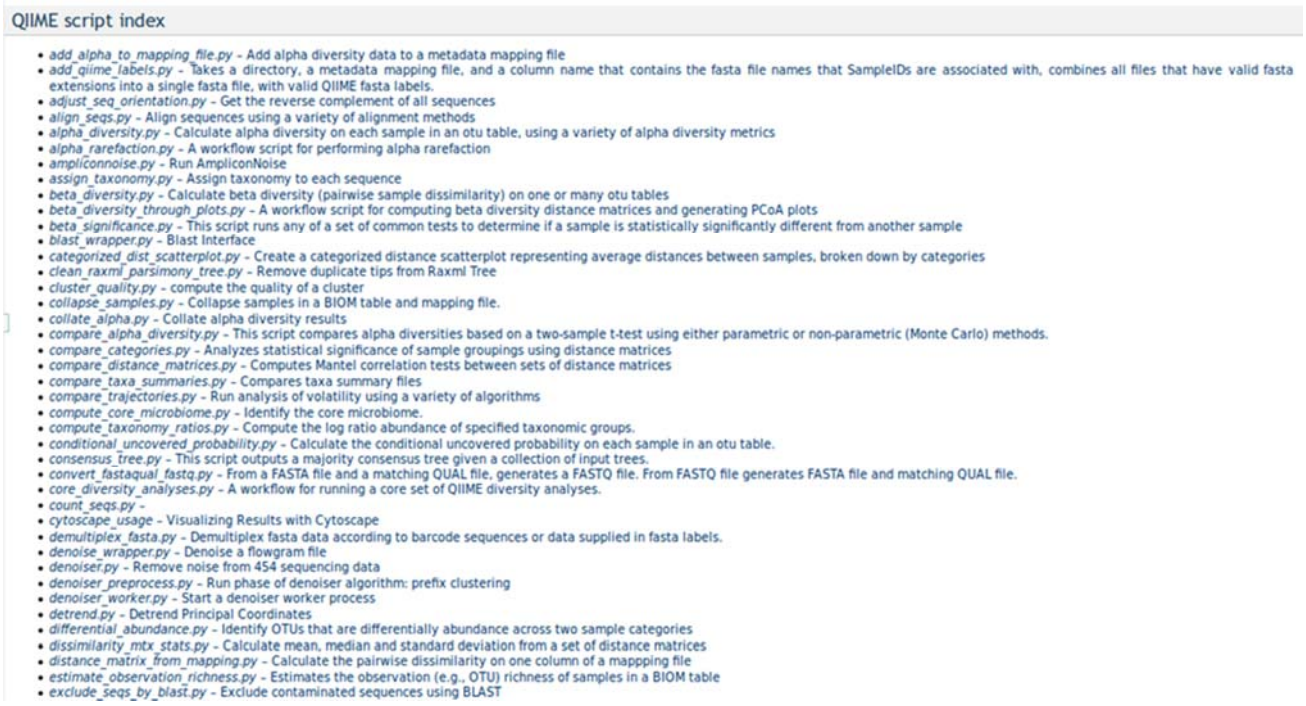
El equipo de La Unidad de Investigación del Hospital Universitario Nuestra Señora de la Candelaria, se decantó por usar QIIME, en lugar de MOTHUR, por varias razones. La principal es que, aunque MOTHUR tiene la gran ventaja de ser bastante más fácil de utilizar e instalar, al contrario que QIIME, con el que hay que lidiar con problemas de librerías y versiones debido al gran número de dependencias y herramienta de terceros que tiene, QIIME provee de una calidad en la visualización de los resultados muy superior a la que obtendríamos si usáramos MOTHUR, lo cual es clave si pensamos que cuanto mejor se puedan visualizar los resultados de los miles o millones de secuencias analizadas, como por ejemplo, en distintos tipos de gráficas o tablas, que nos muestre qué porcentaje de la muestra pertenece a qué especie o qué distribución de individuos existe, si por ejemplo, hacemos un filtrado de la muestra por raza o sexo, en caso de ser de personas o por el tipo de ambiente, si son muestras de agua y aire. Todas estas preguntas se podrán responder mejor o peor en función de cómo uno pueda observar los resultados.

Por tanto, con QIIME, aunque de mayor dificultad de instalación y uso que MOTHUR, disponemos de mejores algoritmos, mayor flexibilidad, una mejor compatibilidad con las plataformas existentes y de una mejor calidad en la observación de los resultados, que de lo que obtendríamos con MOTHUR.

Capítulo 3.
Diseño de la aplicación

3.1 Introducción

QIIME es una herramienta desarrollada en Python que nos permite realizar análisis metagenómicos sobre comunidades microbianas. Es una herramienta muy versátil y poderosa, muy usada en centros de investigación de todo el mundo, capaz de analizar gran cantidad de datos de muestras de muy distinta índole, permitiendo que los investigadores puedan ser más eficientes y rápidos en el desarrollo de sus proyectos. Pero tiene una gran desventaja, y es su gran curva de aprendizaje, que dificulta el ser capaz de poder sacarle partido a todas las posibilidades que ofrece.


QIIME script index

- add_alpha_to_mapping_file.py - Add alpha diversity data to a metadata mapping file
- add_qiime_labels.py - Takes a directory, a metadata mapping file, and a column name that contains the fasta file names that SampleIDs are associated with, combines all files that have valid fasta extensions into a single fasta file, with valid QIIME fasta labels.
- adjust_seq_orientation.py - Get the reverse complement of all sequences
- align_seqs.py - Align sequences using a variety of alignment methods
- alpha_diversity.py - Calculate alpha diversity on each sample in an otu table, using a variety of alpha diversity metrics
- alpha_rarefaction.py - A workflow script for performing alpha rarefaction
- ampliconnoise.py - Run AmpliconNoise
- assign_taxonomy.py - Assign taxonomy to each sequence
- beta_diversity.py - Calculate beta diversity (pairwise sample dissimilarity) on one or many otu tables
- beta_diversity_through_plots.py - A workflow script for computing beta diversity distance matrices and generating PCoA plots
- beta_significance.py - This script runs any of a set of common tests to determine if a sample is statistically significantly different from another sample
- blast_wrapper.py - Blast Interface
- categorized_dist_scatterplot.py - Create a categorized distance scatterplot representing average distances between samples, broken down by categories
- clean_raxml_parsimony_tree.py - Remove duplicate tips from Raxml Tree
- cluster_quality.py - compute the quality of a cluster
- collapse_samples.py - Collapse samples in a BIOM table and mapping file.
- collate_alpha.py - Collate alpha diversity results
- compare_alpha_diversity.py - This script compares alpha diversities based on a two-sample t-test using either parametric or non-parametric (Monte Carlo) methods.
- compare_categories.py - Analyzes statistical significance of sample groupings using distance matrices
- compare_distance_matrices.py - Computes Mantel correlation tests between sets of distance matrices
- compare_taxa_summaries.py - Compares taxa summary files
- compare_trajectories.py - Run analysis of volatility using a variety of algorithms
- compute_core_microbiome.py - Identify the core microbiome.
- compute_taxonomy_ratios.py - Compute the log ratio abundance of specified taxonomic groups.
- conditional_uncovered_probability.py - Calculate the conditional uncovered probability on each sample in an otu table.
- consensus_tree.py - This script outputs a majority consensus tree given a collection of input trees.
- convert_fastaqual_fastq.py - From a FASTA file and a matching QUAL file, generates a FASTQ file. From FASTQ file generates FASTA file and matching QUAL file.
- core_diversity_analyses.py - A workflow for running a core set of QIIME diversity analyses.
- count_seqs.py -
- cytoscape_usage - Visualizing Results with Cytoscape
- demultiplex_fasta.py - Demultiplex fasta data according to barcode sequences or data supplied in fasta labels.
- denoise_wrapper.py - Denoise a flowgram file
- denoiser.py - Remove noise from 454 sequencing data
- denoiser_preprocess.py - Run phase of denoiser algorithm: prefix clustering
- denoiser_worker.py - Start a denoiser worker process
- detrend.py - Detrend Principal Coordinates
- differential_abundance.py - Identify OTUs that are differentially abundance across two sample categories
- dissimilarity_mtx_stats.py - Calculate mean, median and standard deviation from a set of distance matrices
- distance_matrix_from_mapping.py - Calculate the pairwise dissimilarity on one column of a mappping file
- estimate_observation_richness.py - Estimates the observation (e.g., OTU) richness of samples in a BIOM table
- exclude_seqs_by_blast.py - Exclude contaminated sequences using BLAST

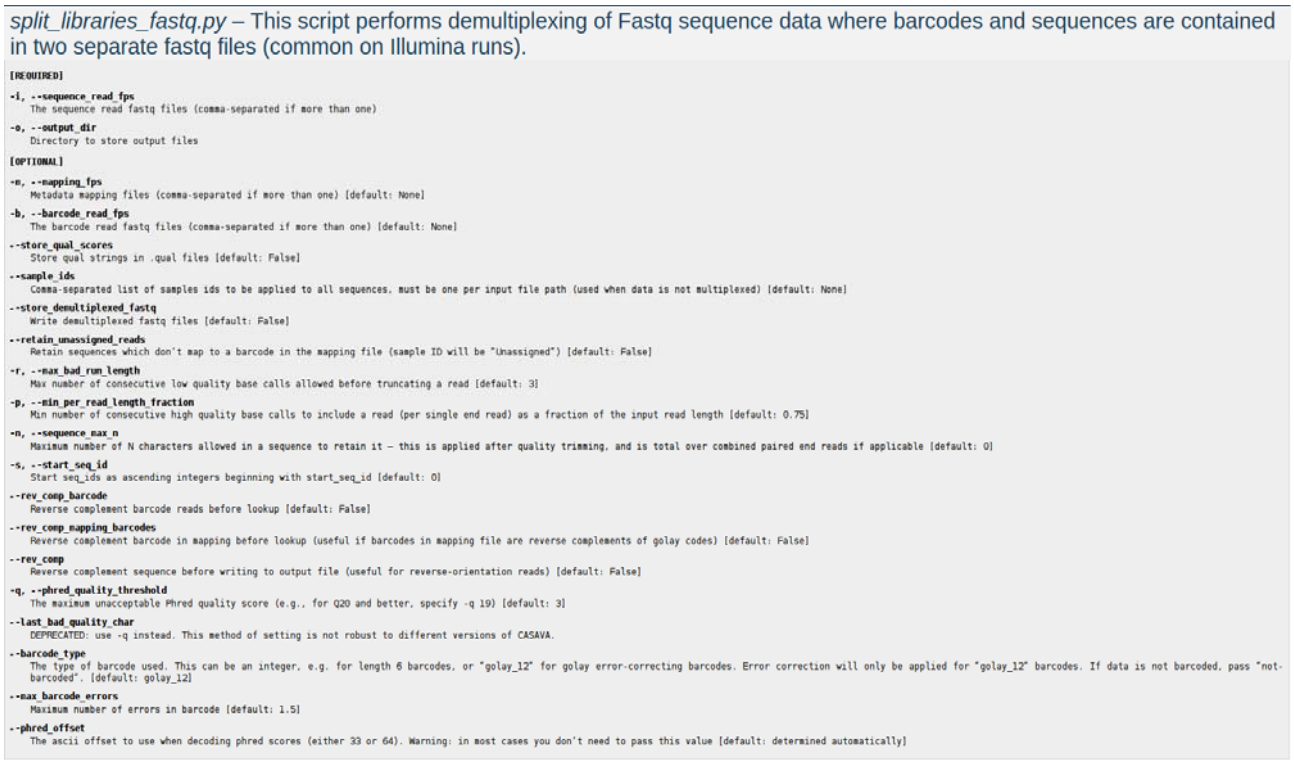
Figura 3.1. Índice con algunos de los scripts incluidos en QUIIME

Es una herramienta que engloba en torno a 150 scripts metagenómicos, todos ellos escritos en Python, en los que cada uno desempeña una función muy concreta. La Figura 3.1 es sólo una pequeña muestra de ellos.

Por defecto, la única forma de interacción que existe con el usuario, es a través de la terminal de comandos de UNIX. Así que para poder utilizarla debemos teclear en el *prompt* de la terminal el nombre del script que deseemos utilizar, seguido de sus correspondientes argumentos. Por lo tanto, debemos utilizar los scripts que nos proporciona QIIME, como si de comandos de UNIX se trataran.

Cada script va acompañado de una serie de argumentos obligatorios y opcionales, todo depende claro, de la tarea que deseemos realizar y los resultados que deseemos obtener. Los argumentos obligatorios son, por ejemplo, indicar al script la ruta del fichero de entrada a tratar o la ruta de salida donde guardar los ficheros resultantes de la ejecución. Los argumentos opcionales por otro lado, por ejemplo, son los ficheros externos que añadan y complementen la información de los ficheros de entrada o parámetros que indican, por ejemplo, umbrales mínimos o máximos en cuanto a la calidad de los datos y poder llevar a cabo un filtrado sobre de las secuencias para eliminar aquellas que cumplan con dicho umbral.

La Figura 3.2 sirve como ejemplo para explicar la dificultad que entraña tener que memorizar cada uno de los scripts que componen QIIME.



split_libraries_fastq.py – This script performs demultiplexing of Fastq sequence data where barcodes and sequences are contained in two separate fastq files (common on Illumina runs).

```
[REQUIRED]
-i, --sequence_read_fps
    The sequence read fastq files (comma-separated if more than one)
-o, --output_dir
    Directory to store output files
[OPTIONAL]
-m, --mapping_fps
    Metadata mapping files (comma-separated if more than one) [default: None]
-b, --barcode_read_fps
    The barcode read fastq files (comma-separated if more than one) [default: None]
--store_qual_scores
    Store qual strings in .qual files [default: False]
--sample_ids
    Comma-separated list of samples ids to be applied to all sequences, must be one per input file path (used when data is not multiplexed) [default: None]
--store_demultiplexed_fastq
    Write demultiplexed fastq files [default: False]
--retain_unassigned_reads
    Retain sequences which don't map to a barcode in the mapping file (sample ID will be "Unassigned") [default: False]
-r, --max_bad_run_length
    Max number of consecutive low quality base calls allowed before truncating a read [default: 3]
-p, --min_per_read_length_fraction
    Min number of consecutive high quality base calls to include a read (per single end read) as a fraction of the input read length [default: 0.75]
-n, --sequence_max_n
    Maximum number of N characters allowed in a sequence to retain it -- this is applied after quality trimming, and is total over combined paired end reads if applicable [default: 0]
-s, --start_seq_id
    Start seq_ids as ascending integers beginning with start_seq_id [default: 0]
--rev_comp_barcode
    Reverse complement barcode reads before lookup [default: False]
--rev_comp_mapping_barcodes
    Reverse complement barcode in mapping before lookup (useful if barcodes in mapping file are reverse complements of golay codes) [default: False]
--rev_comp
    Reverse complement sequence before writing to output file (useful for reverse-orientation reads) [default: False]
-q, --phred_quality_threshold
    The maximum unacceptable Phred quality score (e.g., for Q20 and better, specify -q 19) [default: 3]
--last_bad_quality_char
    DEPRECATED: use -q instead. This method of setting is not robust to different versions of CASAVA.
--barcode_type
    The type of barcode used. This can be an integer, e.g. for length 6 barcodes, or "golay_12" for golay error-correcting barcodes. Error correction will only be applied for "golay_12" barcodes. If data is not barcoded, pass "not-
    barcoded". [default: golay_12]
--max_barcode_errors
    Maximum number of errors in barcode [default: 1.5]
--phred_offset
    The ascii offset to use when decoding phred scores (either 33 or 64). Warning: in most cases you don't need to pass this value [default: determined automatically]
```

Figura 3.2. Ejemplo de la ayuda de un script de QUIIME

Y a continuación, pasaríamos a ejecutar el script para obtener los resultados deseados. En la siguiente imagen de la Figura 3.3 tenemos un ejemplo de cómo deberíamos llamar al script con los argumentos pertinentes. De hecho, en la aplicación se usa la misma cantidad de argumentos para llamar al script.

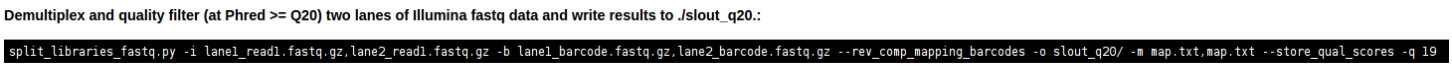


Figura 3.3. Ejemplo de ejecución de un script de QUIIME

Como se puede observar, los scripts funcionan de manera independiente unos de otros, pero lo cierto es que la salida resultante de uno se convierte en el fichero de entrada para el siguiente, que debemos por supuesto, indicar en la cola de argumentos, y así continuamente con todos aquellos scripts que queramos utilizar.

Por tanto, en este punto podemos destacar, por un lado, la gran flexibilidad que nos aporta QIIME, que, al ser una herramienta completamente modular, nos permite, utilizar sólo aquellos scripts que necesitemos. Pero al mismo tiempo observamos, que esa misma flexibilidad e independencia que nos aporta, deja al usuario solo, ante la tarea de tener que desarrollar sus propias interfaces o *workflows* para poder trabajar de la manera más cómoda y eficiente posible, claro está, una vez que haya sido capaz de asimilar y comprender por completo el funcionamiento de QIIME. Una vez alcance esa posición, será capaz de trabajar de manera completamente autónoma con QIIME y ser capaz, como es también nuestro caso, complementarla con otras herramientas externas.

Cabe destacar, que a pesar de la enorme dificultad que supone para cualquier usuario, utilizar de manera óptima QIIME y de lidiar con todos los problemas que se pueden ir presentando, se añade un grado de dificultad más si cabe. Y es que como en el caso del personal de la Unidad de Investigación del Hospital y de la Unidad de Genómica Aplicada de la Universidad, no son usuarios habituados a trabajar a través de la línea de comandos, que es por defecto la interfaz por la que se debe trabajar con QIIME, por lo que requieren de ayuda externa para poder realizar todas estas labores, papel que está pensado que sea desempeñado por los bioinformáticos.

3.2 Diseño de la aplicación

Partiendo de la premisa anterior, la labor encomendada fue la de desarrollar una aplicación que permitiera al usuario poder trabajar con QIIME, pero sin tener que comprender cada uno de sus entresijos, y poder interactuar directamente con él, es decir, que para él le fuese totalmente transparente, que no necesitase conocer cada uno de los scripts y cómo funcionan, para poder desarrollar su trabajo. El diseño de dicha herramienta debería poder llevar al usuario desde unos sencillos pasos iniciales, como, por ejemplo, indicar cuáles son los ficheros que quiere analizar y definir ciertos parámetros para el análisis, más que nada para poder adaptar la aplicación a según qué proyectos, hasta el resultado final, y que el usuario pueda visualizar e interactuar con los resultados de los análisis llevados a cabo internamente por QIIME.

El diseño de la aplicación tendría dos partes bien definidas. Por un lado, se desarrollaría un **pipeline** que aglutinase todos aquellos scripts que requerían, en este caso en concreto, el personal investigador de la Unidad de Investigación del Hospital y Unidad de Genómica Aplicada de la universidad. Una vez hecho esto, el pipeline se desarrollaría de manera que los scripts se fuesen ejecutando de manera autónoma, siguiendo un estricto flujo de ejecución. Internamente, el pipeline también se encargaría de todo lo concerniente con el manejo de los ficheros y directorios que se irían creando conforme va avanzando la ejecución y se encargaría de pasárselos como ficheros de entrada a los scripts que fuese necesarios. Además, el pipeline tendría que tener la suficiente disponibilidad para poder analizar datos tanto de Illumina como de Ion Torrent.

Por otro lado, se desarrollaría una **capa web** que englobase a su vez al pipeline. La parte web sería la encargada de la interacción directa con el usuario. A través de una serie de formularios web, le iría pidiendo al usuario, cuáles son los ficheros que desea analizar y cuáles son los parámetros con los que quiere que sea llevado a cabo el análisis. Una vez hayan sido cumplimentados todos los formularios, procederíamos a pasarle los ficheros y parámetros designados por el usuario, al pipeline. A partir de este punto, tomaría el control y se encargaría de ejecutar QIIME, tratar los datos debidamente, llevar a cabo los análisis y devolver los resultados a la capa

web, para que ésta se los presente al usuario y pueda visualizarlos e pueda interactuar con ellos.

Se ha elegido hacerlo vía web en lugar de desarrollar una aplicación de escritorio por varias razones. La primera, que permite poder acceder desde múltiples terminales a la aplicación y trabajar con ella, en caso de que la aplicación sea instalada, por ejemplo, en un servidor. La segunda, es porqué haciéndolo vía web, nos ahorramos temas de compatibilidades con múltiples plataformas, la necesidad de tener que instalar programas y/o librerías adicionales para que todo funcione correctamente. Así que, de esta manera, sólo debemos tener instalados QIIME y las herramientas adicionales que queramos incluir en el pipeline, y el servidor web encargado de gestionar la capa web y las peticiones de los usuarios.

El diseño quedaría tal y como se indica en el siguiente esquema de la Figura 3.4.

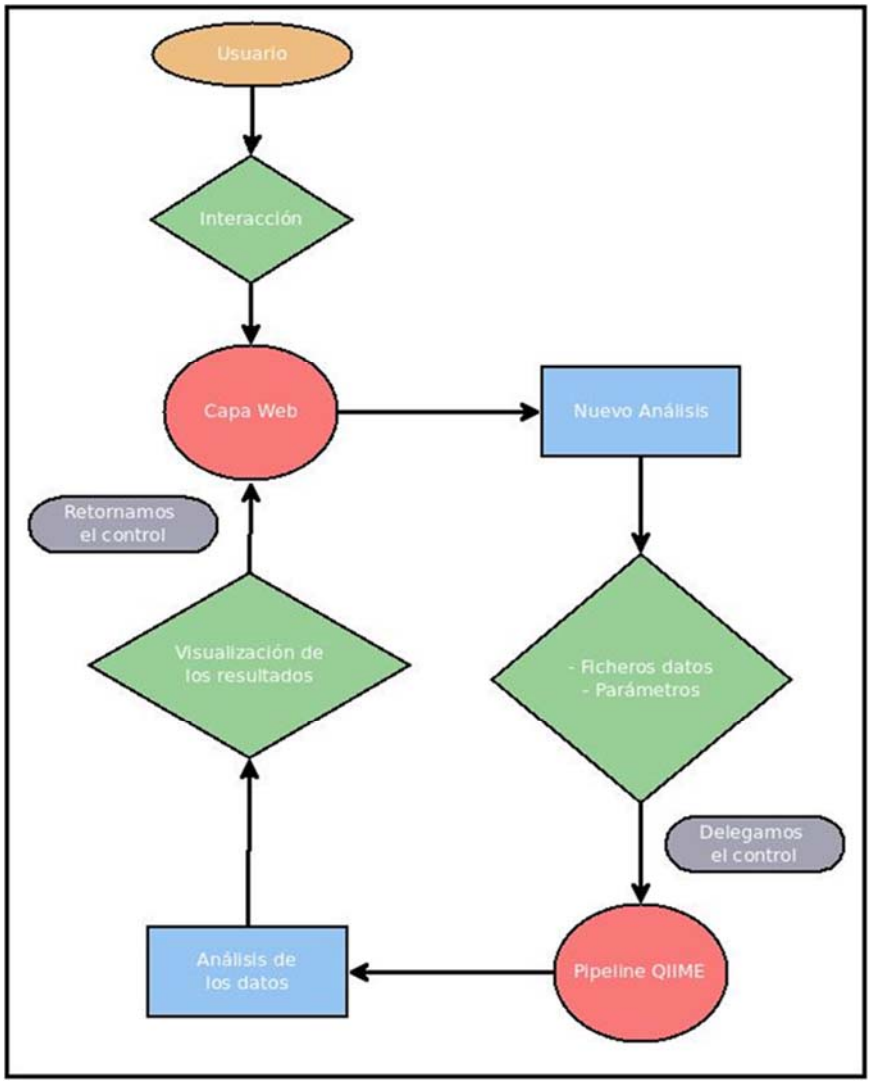


Figura 3.4. Esquema del diseño de la aplicación

Capítulo 4.
Desarrollo de la aplicación

4.1 Desarrollo del *pipeline*

Tal y como se ha comentado en secciones anteriores, QIIME es una herramienta con la que se interactúa a través de la línea de comandos. Teniendo esto en mente, decidimos que debíamos desarrollar el pipeline de trabajo sobre el mismo entorno, la interfaz de línea de comandos. Por tanto, la solución final que se adoptó fue la de implementarlo en Bash, el lenguaje de programación de consola de UNIX. Bash, además de ser una de las shell de UNIX más ampliamente utilizada, es un lenguaje de programación muy potente, que nos permite utilizar, por un lado, la potencia de todos los comandos que existen en el mundo UNIX, junto con el uso de las características básicas de las que disponen todos los lenguajes de programación modernos, como sentencias de control de flujo, implementación de subrutinas, etc.

La estrategia que seguimos de cara a plantear el diseño y estructura que debería seguir el pipeline, nos la marcó la propia herramienta, QIIME. Debíamos, antes que nada, hacernos una idea de todos los pasos que había que seguir, para llegar desde los pasos iniciales, donde se realiza un tratamiento previo de datos en crudo, hasta el paso final, donde poder llevar a cabo el análisis de esos datos y que el usuario pudiera visualizar los resultados obtenidos. En resumen, los pasos que teníamos que seguir eran los siguientes:

- Validar los datos recibidos por parte del usuario y prepararlos para los pasos posteriores.
- Clasificar las muestras según un criterio específico dado.
- Realizar un filtrado de muestras no deseadas sobre los resultados de la clasificación anterior. Así podremos verificar que no existen muestras contaminadas, muestras de control o datos obtenidos con demasiada

incertidumbre. Este paso nos ayudará a conseguir unos resultados óptimos y los más claros posibles para su investigación. Es un proceso obligado al ser las muestras recogidas de entornos donde conviven multitud de especies diferentes.

- Con las muestras ya correctamente clasificadas y limpias, proceder a su análisis.
- Mostrar los resultados obtenidos a través de la interfaz web al usuario, para que éste pueda tomar las conclusiones oportunas.

Una vez claros los pasos generales que debíamos seguir para conseguir nuestros objetivos, pasamos al trabajo con QIIME. Primero que nada, debíamos conoces cuáles de todos los scripts que conforman la herramienta, nos eran realmente necesarios en nuestra tarea. Con ayuda de la documentación de QIIME, se hizo un listado de todos los scripts involucrados en el proceso.

También hubo que tener claro cuáles serían los procesos intermedios que habría que hacer entre la ejecución de un script al siguiente. Es decir, si debíamos, por ejemplo, editar determinados ficheros eliminando la información considerada como no relevante para el propósito que perseguíamos y así, sólo quedarnos con aquella que nos es realmente útil. Otra tarea intermedia, sería todo lo relacionado con la correcta gestión de los ficheros y directorios que se irían creando a medida que avanzaba la ejecución del pipeline.

Cuando ya se tuvo claro todo lo había que tener en cuenta si queríamos que todo funcionase correctamente, se decidió que el mejor diseño posible para el pipeline, era dividirlo en varias partes, es decir, hacer un diseño modular y separar en diferentes scripts las funcionalidades clave. Estructurando de esta manera el pipeline, se conseguía, por un lado, facilitar la tarea a la hora de tener que depurar errores y realizar el debido mantenimiento a posteriori, y, por otro lado, nos facilitaba la tarea a la hora de añadir futuras modificaciones o añadir nuevas herramientas que pudiesen complementar de lo que ahora ya disponíamos.

El diseño final del pipeline principal lo conforman cinco scripts, tal y como queda reflejado en este esquema de la Figura 4.1:

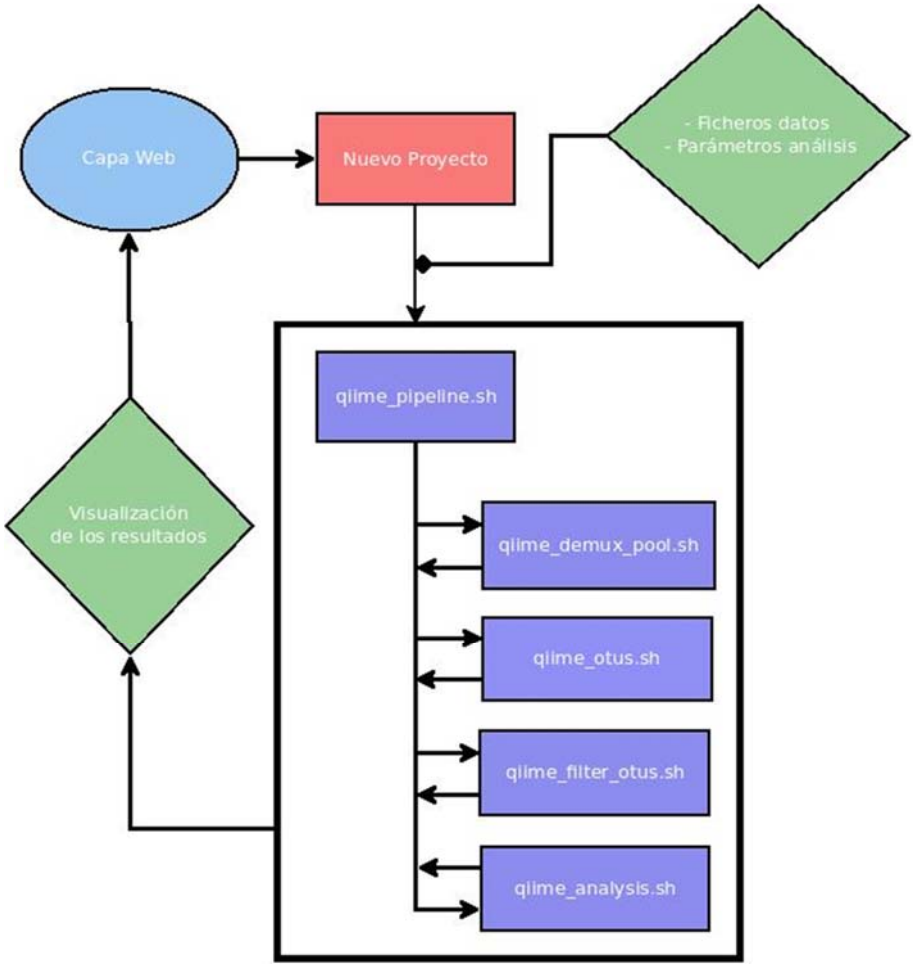


Figura 4.1. Esquema del diseño del pipeline

4.1.1 Descripción de los scripts

Como podemos observar en el esquema anterior, la ejecución del pipeline, parte del script **qiime_pipeline.sh**, que, a su vez, es el que engloba la ejecución de todos los demás. A él es, al que la aplicación web llamará y le pasará los ficheros de datos y los parámetros de configuración del análisis, y a partir de ahí, irá ejecutando el resto de los componentes del pipeline secuencialmente, gestionando los resultados que se vayan obteniendo en cada paso del proceso hasta culminar en la visualización de los resultados, por parte de la aplicación web.

A continuación, se explicarán en detalle la funcionalidad que desempeña cada uno dentro del pipeline.

qiime_pipeline.sh

Es el script encargado de gestionar todo el trabajo que se desarrolla en el pipeline. Este es el script, al que llamamos desde la aplicación web y al que le pasamos los ficheros de datos y los parámetros definidos por el usuario para desarrollar el trabajo requerido.

Este script tiene como primera funcionalidad validar uno de los ficheros del usuario, el fichero **mapping file**. Es un fichero de texto plano, que contiene una tabla con información acerca de las secuencias, es decir, es un fichero con metadatos sobre cada una de las muestras. Información, por ejemplo, sobre el individuo al que pertenece dicha muestra, con datos sobre él como el sexo o la raza.

Sólo puede existir un **mapping file** por pool de datos. Un ***pool* de datos**, es un conjunto de muestras independientes que han sido secuenciadas en un experimento. La información de las secuencias a analizar estará contenida en estos pools. En caso de que los datos hayan sido secuenciados a través de Illumina, serán tres los ficheros que compondrán un pool de datos, un fichero *forward*, un fichero de *reverse* y un tercero, el de índices o *barcodes*. Se denominan así, porque en Illumina, se secuencian las cadenas de ADN desde el inicio y el final de la cadena de manera simultánea, lo que se denomina como *paired-end reads*. El fichero de barcodes contiene indicadores de correspondencia de cada secuencia a cada muestra.

En caso de ser de un secuenciador de IonTorrent, serán sólo dos, el fichero con los datos de la secuencia y el de barcodes con los indicadores de calidad, lo que se denomina como *single-end read.*

Hay que destacar que se pueden analizar múltiples pools de datos en una sesión, cada debe tener todos sus ficheros correspondientes.

qiime_demux_pool.sh

Este script tiene como función demultiplexar las secuencias. Como ficheros de entrada, tendrá el fichero mapping y el pool de datos. Los ficheros que contienen los pools de datos, son ficheros de tipo **FASTQ**, un formato estandarizado basado en texto, para almacenar las secuencias.

Seguidamente, debemos unir las lecturas en un único archivo, aunque esto sólo debemos hacerlo para el caso de Illumina, que unirá los dos ficheros FASTQ, el forward y el reverse en un único archivo.

A continuación, procederemos a la demuliplexación del fichero resultante del paso anterior. Este paso consiste en dividir el fichero tipo FASTQ en otros dos ficheros distintos, por un lado, un fichero tipo **FASTA** o **FNA**, que contiene los datos de las secuencias y por otro lado un fichero tipo **QUAL** que contendrá los indicadores de calidad para cada secuencia. Tal y como se muestra en la siguiente Figura 4.2:

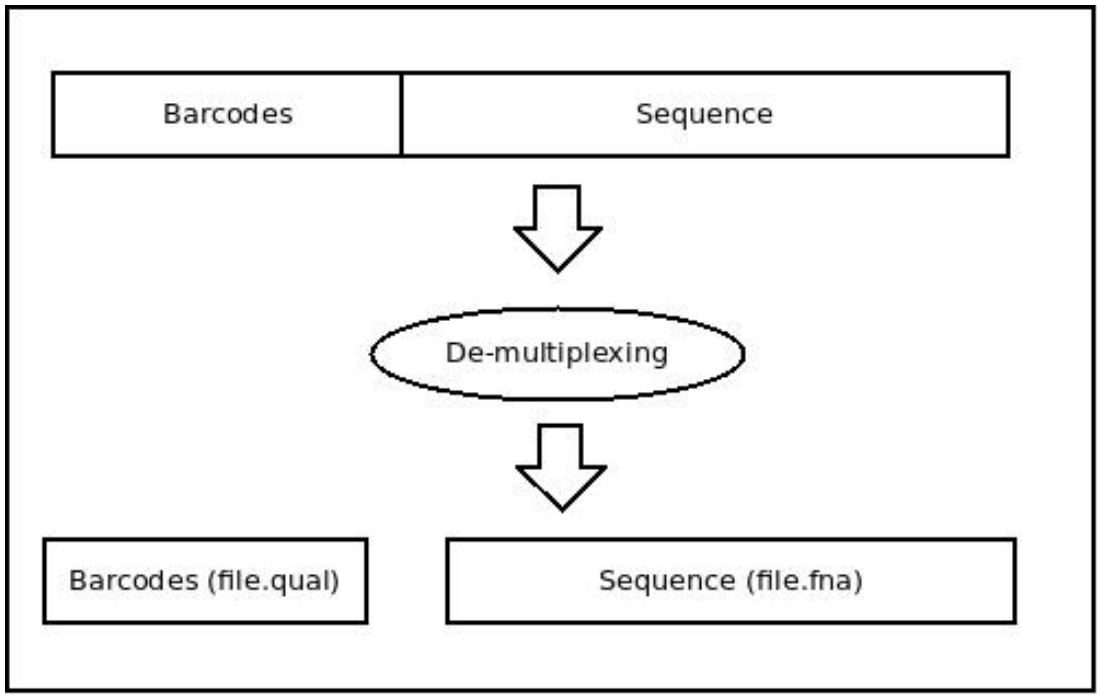


Figura 4.2. Proceso de demultiplexación

Es precisamente ese fichero FNA el que necesitaremos para los procesos posteriores. Los scripts de QIIME involucrados en este paso son:

- Para el caso de "paired-ends read" (caso de Illumina)
  - join_paired_ends.py (unir ficheros forward y reverse)
  - split_libraries_fastq.py (demultiplexar)
- Para el caso de "single-end read" (IonTorrent principalmente)
  - convert_fastaqual_fastq.py
  - split_libraries.py

qiime_otu.sh

Tiene como función, construir la tabla de **OTUs** (*Operational Taxonomy Units*) [12] y el árbol de filogenia. A partir de este punto en adelante sólo se trabajará con las denominadas tablas OTUs, por ello, éste es un paso muy importante dentro del pipeline, además de resultar el más costoso en términos de consumo de tiempo y recursos de cómputo.

Como ficheros de entrada principal tenemos el fichero tipo FNA resultante en el paso anterior. Las OTUs son cluster o agrupaciones de secuencias que guardan cierto de grado de similitud entre sí. Por tanto, permite clasificar las secuencias por la taxonomía de los individuos de la muestra. Es a partir de la tabla de OTUs resultante, con la que podremos llevar a cabo los análisis.

Para llevar a cabo este proceso de clustering, QIIME utiliza por defecto el algoritmo **UCLUST** [13]. Este proceso en QIIME se denomina *OTU picking* y existen otros algoritmos con los que poder construir las tablas de OTUs.

UCLUST es un algoritmo publicado en el año 2010, capaz de realizar clustering, alineamiento y búsqueda sobre millones de secuencias. El proceso que nos concierne en este proyecto es el clustering para secuencias 16S rRNA.

Para el proceso de clustering, UCLUST usa una base de datos de referencia con un grado de similitud del 97% (por consenso, a nivel de secuencia, una especie diferirá de otra como media en un 3% de la longitud comparada), y va comparando las secuencias de entrada con la base de datos. Si tienen similitud, es añadida al cluster correspondientes. En caso de no haber similitud, se crea un nuevo cluster y se añade a él.

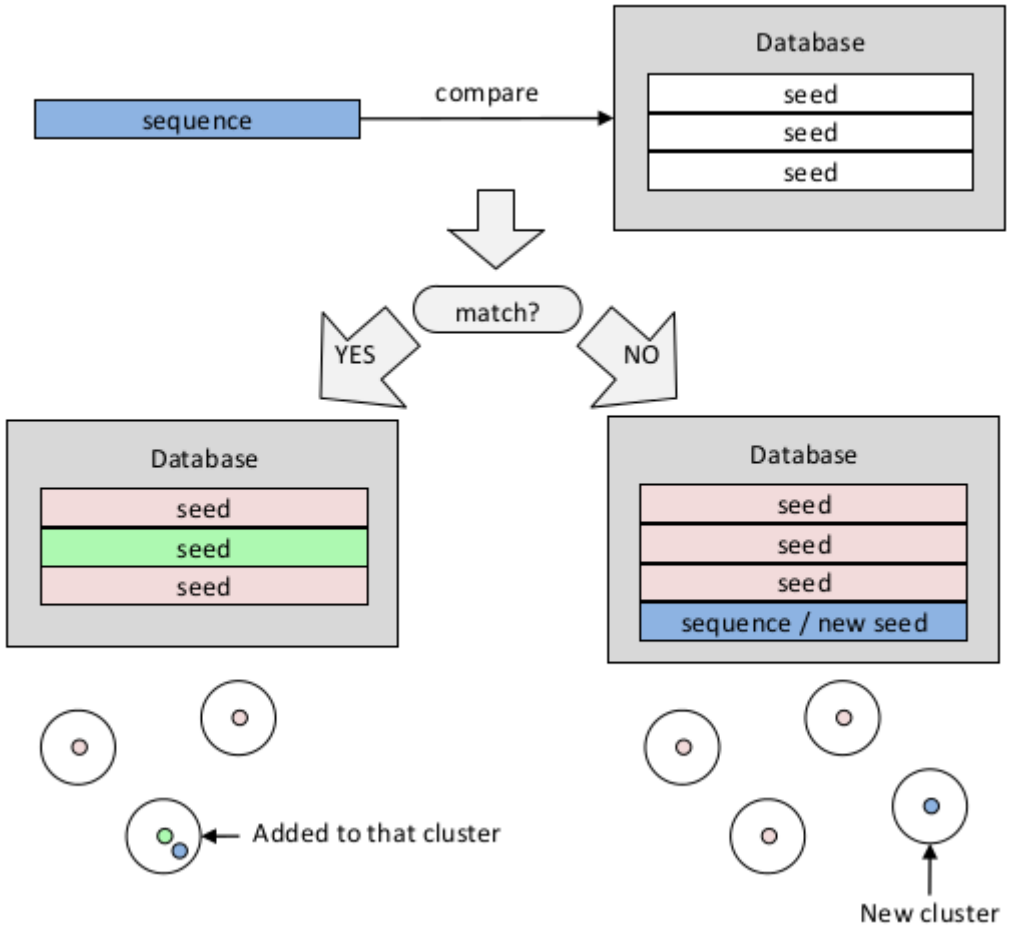


Figura 4.3. Proceso de clustering

Como resultado, obtendremos una tabla de OTUs y el árbol de filogenia. El árbol de filogenia es una representación gráfica en forma de árbol.

Los scripts de QIIME involucrados en este paso son:

- pick_open_reference_otus.py (Paso más costoso)
- parallel_identify_chimeric_seqs.py
- filter_fasta.py
- assign_taxonomy.py
- make_otu_table.py
- parallel_align_seqs_pynast.py
- filter_alignment.py
- make_phylogeny.py

qiime_filter_otus.sh

Este script tiene la función de limpiar la tabla de OTUs resultante del proceso anterior. En este proceso de filtrado, lo que se busca es eliminar de la tabla las muestras contaminantes, las muestras que hayan sido marcadas como muestras de control, o las muestras que no se ciñen a la taxonomía deseada, es decir, que sean de individuos ajenos a los que queremos estudiar. El archivo mapping, contiene la información, sobre qué muestras del total, son las del control. Por ello es uno de los ficheros de entrada, junto con la tabla de OTUs.

Los scripts de QIIME involucrados en este paso son:

- filter_taxa_from_otu_table.py (filtrado por taxonomía)
- filter_samples_from_otu_table.py (filtrar por muestras de control)
- filter_otus_from_otu_table.py (filtrar las muestras contaminantes)

qiime_analysis.sh

Es el encargado de llevar a cabo los análisis estadísticos sobre la tabla de OTUs. Una vez tenemos una tabla final de OTUs limpia con la que poder trabajar, es hora de lanzar los análisis sobre ella y visualizar los resultados obtenidos. Los análisis nos arrojarán algo de luz sobre los individuos que residen en la muestra y en qué cantidad lo hacen para cada una. Además, podremos ver la cantidad de individuos por el tipo de muestra, es decir, en nuestro caso, que porcentaje de la bacteria X, está presente en personas de la raza Y y de sexo Z.

El script utilizado en este paso es el **core_analysis.py**. El script en sí es un pequeño pipeline, que engloba otros scripts para análisis de diversidad como son:

- alpha_rarefaction.py
- beta_diversity_thorugh_plots.py
- summarize_taxa_through_plots.py

4.1.2 Herramientas adicionales: PICRUST y FastQC

Una vez concluida la parte principal, la concerniente a QIIME, vamos a describir dos herramientas adicionales que se han decidido incluir en el pipeline para complementar los resultados de QIIME con información adicional. Las herramientas en cuestión son **PICRUST** [**14**] y **FastQC** [**15**].

PICRUST es una herramienta, que, a partir de una población de individuos de una comunidad microbiana, es capaz de predecir las funciones que se realizan en dicha comunidad. Es decir, para PICRUST, no es tan importante quién vive en una comunidad, como lo que son capaces de desempeñar en dicha comunidad.

PICRUST funciona perfectamente junto a QIIME, porqué de hecho, es con los resultados de las tablas de OTUs (la tabla con la información de los individuos), con los que realiza sus predicciones sobre las funciones celulares llevadas a cabo en la comunidad. Luego una vez tenemos dichas predicciones, volvemos a realizar los análisis estadísticos de QIIME sobre dichas predicciones, lo que nos permitirá tener una idea más clara sobre que procesos celulares ha detectado PICRUST, y lo que es más importante aún, el porcentaje que existe de ellos en cada muestra del experimento.

Por otro lado, tenemos FastQC, es una herramienta, desarrollada en Java, para la visualización de las calidades y longitudes asociadas a las lecturas (secuencias) de un experimento. El caso que nos atañe, la usaremos para poder visualizar el antes y el después de las muestras, es decir el antes del filtrado y después del mismo.

Como resultado, nuestro esquema del pipeline, al añadir estas dos nuevas herramientas, quedaría como representa la siguiente Figura 4.4:

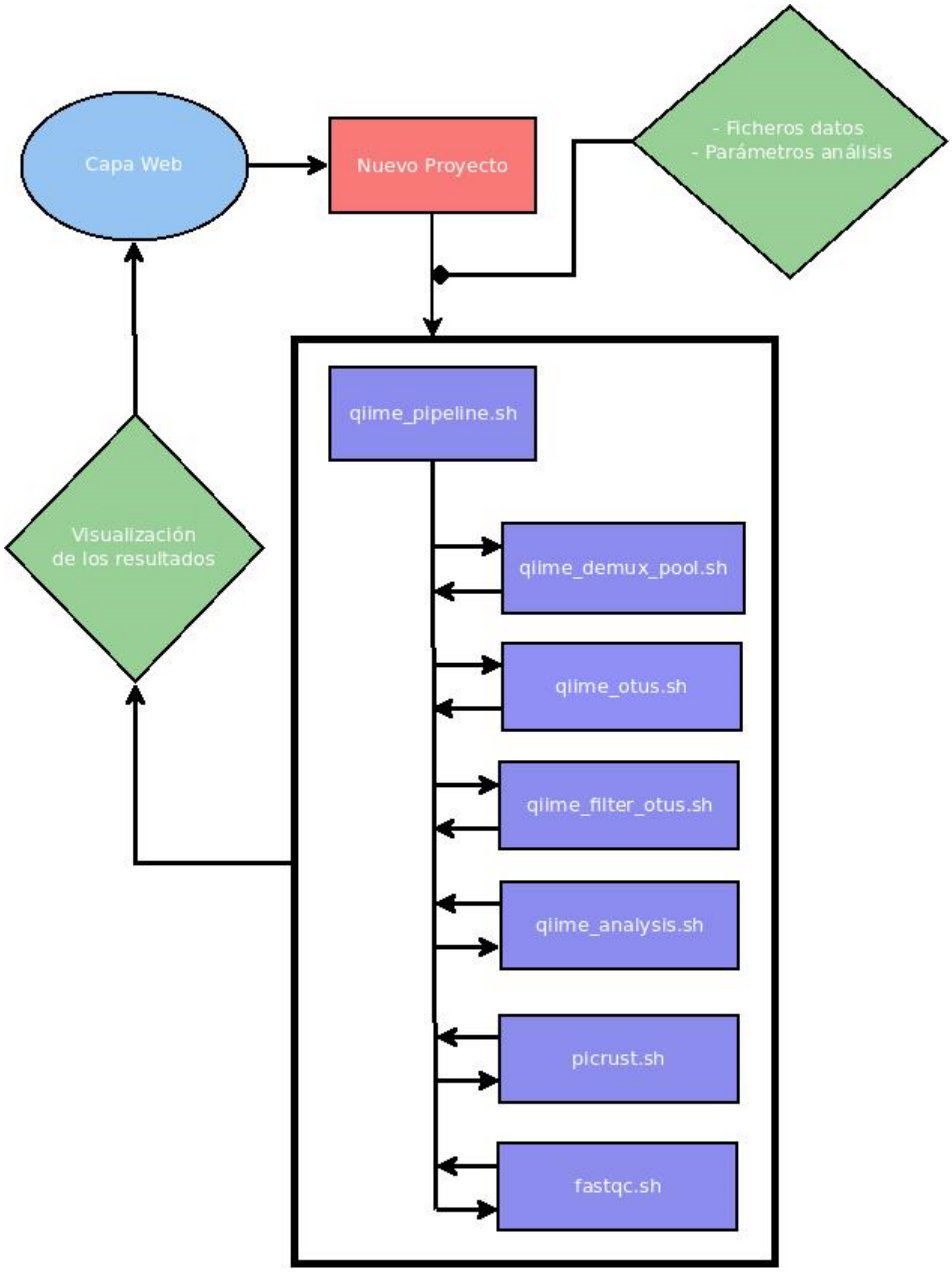


Figura 4.4. Esquema de diseño del nuevo pipeline

4.2 Desarrollo de la aplicación Web

La capa web del proyecto, es una aplicación web, que ha sido desarrollada utilizando el *framework* de desarrollo web **Django** [16]. Es un *framework* de desarrollo web que se ha ido consolidando a lo largo de los años, desde que se publicase su primera versión allá por el año 2005. Gracias a la enorme comunidad que existe a su alrededor, provee al desarrollador de multitud de módulos para implementar cualquier funcionalidad que requiera una aplicación web moderna.

Django nos permite desarrollar completas aplicaciones web, en espacios de tiempo muy cortos y sin invertir excesivos recursos, lo cual es perfecto para la naturaleza de este proyecto, en el que la mayor parte del peso recae sobre el pipeline anteriormente mencionado. Otra de las razones por la que se eligió este *framework* y no otro, es porque está escrito enteramente en Python, y todo lo que tengamos que desarrollar con él, será desarrollado usando justamente el mismo lenguaje de programación. Al necesitar de una interacción directa con el sistema operativo, Python posee, mejor que otros lenguajes, módulos e interfaces para realizar llamadas al sistema e interactuar con el sistema operativo de la manera más fácil y cómoda posible. Por ello, Python es un lenguaje muy demandado en el ámbito de la administración de sistemas y redes de computadores.

Además, Python es el lenguaje usado por excelencia por el mundo científico, con el que ha sido posible implementar soluciones de software a problemas presentados en campos tan diversos como física, astronomía, química o biología, que es el campo que ahora nos concierne. Hasta el punto de que es posible encontrar programas y paquetes específicos para cada rama de la ciencia. No es por ello de extrañar, que QIIME y la inmensa mayoría de las herramientas bioinformáticas disponibles en el mercado estén siendo desarrolladas usando en este lenguaje.

La aplicación web consta de una serie de formularios donde el usuario, escogerá primero, que tipo de proyecto desea crear, si es un proyecto de Illumina o uno de IonTorrent, seleccionará los ficheros de datos que desea que sean analizados y elegirá la configuración de trabajo para el Pipeline que se adapte al tipo de resultados que desea obtener. Una vez elegidos los ficheros y la configuración de trabajo, al usuario se le mostrará un resumen detallado de lo que ha elegido y con lo que el pipeline trabajará. A partir de aquí empezará a trabajar el pipeline.

Una vez haya concluido el análisis por completo, al usuario se le mostrará la pantalla de resultados, en la que podrá visualizar e interactuar con los resultados obtenidos de su análisis.

Capítulo 5. Resultados

5.1 Introducción

En la pantalla de resultados de la aplicación, el usuario podrá acceder a la información de los resultados de su análisis, que se encuentra divida en cuatro partes. Por un lado, podemos encontrar la información concerniente a las características del proyecto que acaba de ser ejecutado. Datos como el tipo de proyecto, la ruta donde se encuentran todos los ficheros resultantes de la ejecución del pipeline, e información relevante en cuanto al rendimiento del análisis como el tiempo total que ha tardado en ejecutar el proceso. Estos datos, junto con los resultados propios del análisis son muy importantes, puesto que abren la puerta a la posibilidad de realizar estudios comparativos entre análisis llevados a cabo sobre una misma muestra, pero con parámetros distintos, como tipo de secuenciador utilizado, Illumina o IonTorrent, estrategia escogida a la hora de construir de las tablas de OTUs, rendimiento de la propia máquina donde se lleva a cabo el trabajo, etc. Abre un abanico de posibilidades muy amplio. De hecho, ya existen muchos estudios [17], que llevan a cabo análisis comparativos, sobre la misma muestra, usando varias herramientas bioinformáticas distintas, para medir el rendimiento de cada una y la eficacia del análisis que llevan a cabo.

Por otra parte, tenemos los resultados obtenidos del análisis de QIIME, los de PICRUST y los de FastQC, a todos ellos podremos acceder a través de un enlace. Los resultados a los que se accede, está conformado en su mayor parte por una serie de gráficas, histogramas, gráficas de áreas y gráficas en 3D, con la que se puede tener una visión clara de los resultados obtenidos.

Los resultados que se van a mostrar están realizados con secuencias de 250 pares de bases PE de Illumina en lecturas de la región V4 en muestras de ADN de saliva de niños de diverso origen étnico tomadas como parte del estudio GALA II.

5.2 Pantalla de resultados

En la Figura 5.1, un usuario puede observar los resultados de la aplicación web, con datos técnicos del proyecto, como el tipo, la ruta donde se aloja el directorio el proyecto, el tiempo tardado en ejecutarlo, etc. También tenemos información como el número de secuencias por pool de datos que han sido analizadas, el número total de OTUs encontradas. Además, al final de la pantalla de resultados, se representan gráficamente los histogramas (Figura 5.2), un histograma por pool de datos con las lecturas de cada uno. Y por último a la derecha encontraremos los enlaces hacia los resultados de QIIME, PICRUST y FastQC.

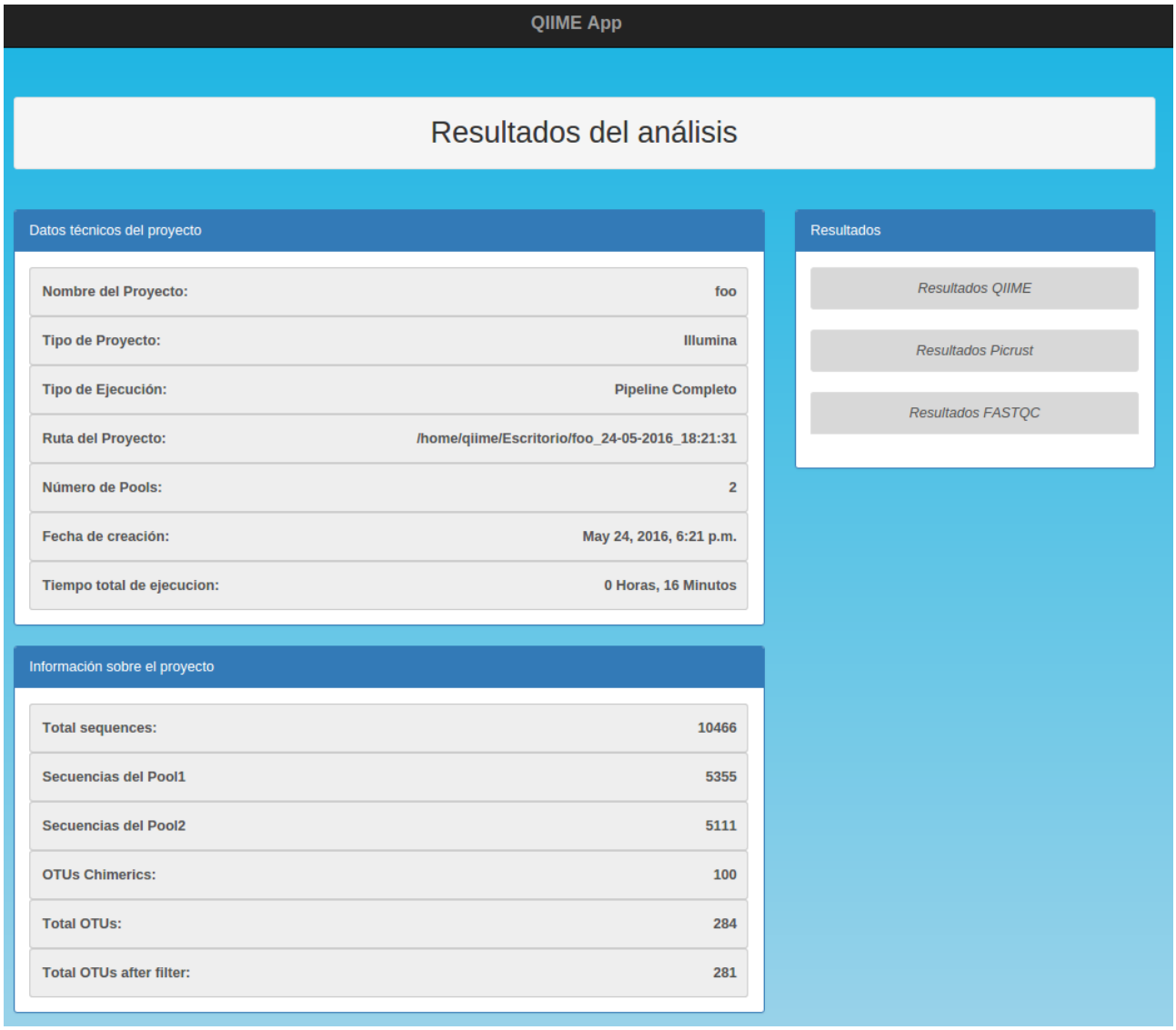


Figura 5.1. Pantalla de resultados

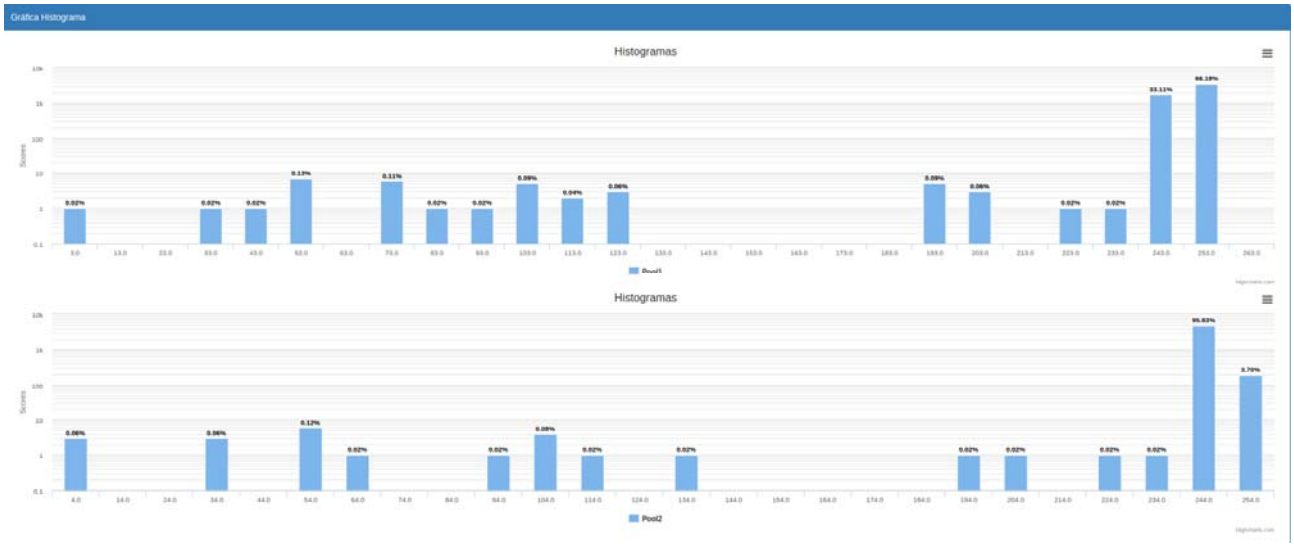


Figura 5.2. Pantalla de histogramas

Run summary data	
Master run log	log_20160226155111.txt
BIOM table statistics	biom_table_summary.txt
Filtered BIOM table (minimum sequence count: 100)	table_mc100.biom.gz
rarefied BIOM table (sampling depth: 100)	table_even100.biom.gz
Taxonomic summary results (by Ethnicity)	
Taxa summary bar plots	bar_charts.html
Taxa summary area plots	area_charts.html
Taxonomic summary results (by Trait)	
Taxa summary bar plots	bar_charts.html
Taxa summary area plots	area_charts.html
Beta diversity results (even sampling: 100)	
Distance boxplots (weighted_unifrac)	Trait_Distances.pdf
Distance boxplots statistics (weighted_unifrac)	Trait_Stats.txt
Distance boxplots (weighted_unifrac)	Ancestry_Distances.pdf
Distance boxplots statistics (weighted_unifrac)	Ancestry_Stats.txt
Distance boxplots (weighted_unifrac)	Ethnicity_Distances.pdf
Distance boxplots statistics (weighted_unifrac)	Ethnicity_Stats.txt
Distance boxplots (weighted_unifrac)	Gender_Distances.pdf
Distance boxplots statistics (weighted_unifrac)	Gender_Stats.txt
PCoA plot (weighted_unifrac)	index.html
Distance matrix (weighted_unifrac)	weighted_unifrac_dm.txt
Principal coordinate matrix (weighted_unifrac)	weighted_unifrac_pc.txt
Distance boxplots (unweighted_unifrac)	Trait_Distances.pdf
Distance boxplots statistics (unweighted_unifrac)	Trait_Stats.txt
Distance boxplots (unweighted_unifrac)	Ancestry_Distances.pdf
Distance boxplots statistics (unweighted_unifrac)	Ancestry_Stats.txt
Distance boxplots (unweighted_unifrac)	Ethnicity_Distances.pdf
Distance boxplots statistics (unweighted_unifrac)	Ethnicity_Stats.txt
Distance boxplots (unweighted_unifrac)	Gender_Distances.pdf
Distance boxplots statistics (unweighted_unifrac)	Gender_Stats.txt
PCoA plot (unweighted_unifrac)	index.html
Distance matrix (unweighted_unifrac)	unweighted_unifrac_dm.txt
Principal coordinate matrix (unweighted_unifrac)	unweighted_unifrac_pc.txt
Taxonomic summary results	
Taxa summary bar plots	bar_charts.html
Taxa summary area plots	area_charts.html
Taxonomic summary results (by Gender)	
Taxa summary bar plots	bar_charts.html
Taxa summary area plots	area_charts.html
Alpha diversity results	
Alpha rarefaction plots	rarefaction_plots.html

Figura 5.3. Datos de resumen de QUIIME

5.2.1 Resultados de QIIME

Los resultados de QIIME, aparecen englobados en una tabla autogenerada por el propio programa, que, dividida en diferentes secciones, nos proporciona toda la información sobre los resultados de los análisis llevados a cabo. Podremos visualizar los resultados de diferentes maneras (ver Figura 5.3).

La Figura 5.4 es un ejemplo de cómo, mediante una gráfica de barras, se nos representa la cantidad de individuos que hay por muestra. Debajo de la gráfica (Figura 5.5) tenemos un cuadro informativo sobre qué individuos son y el porcentaje que existe de ellos por muestra.

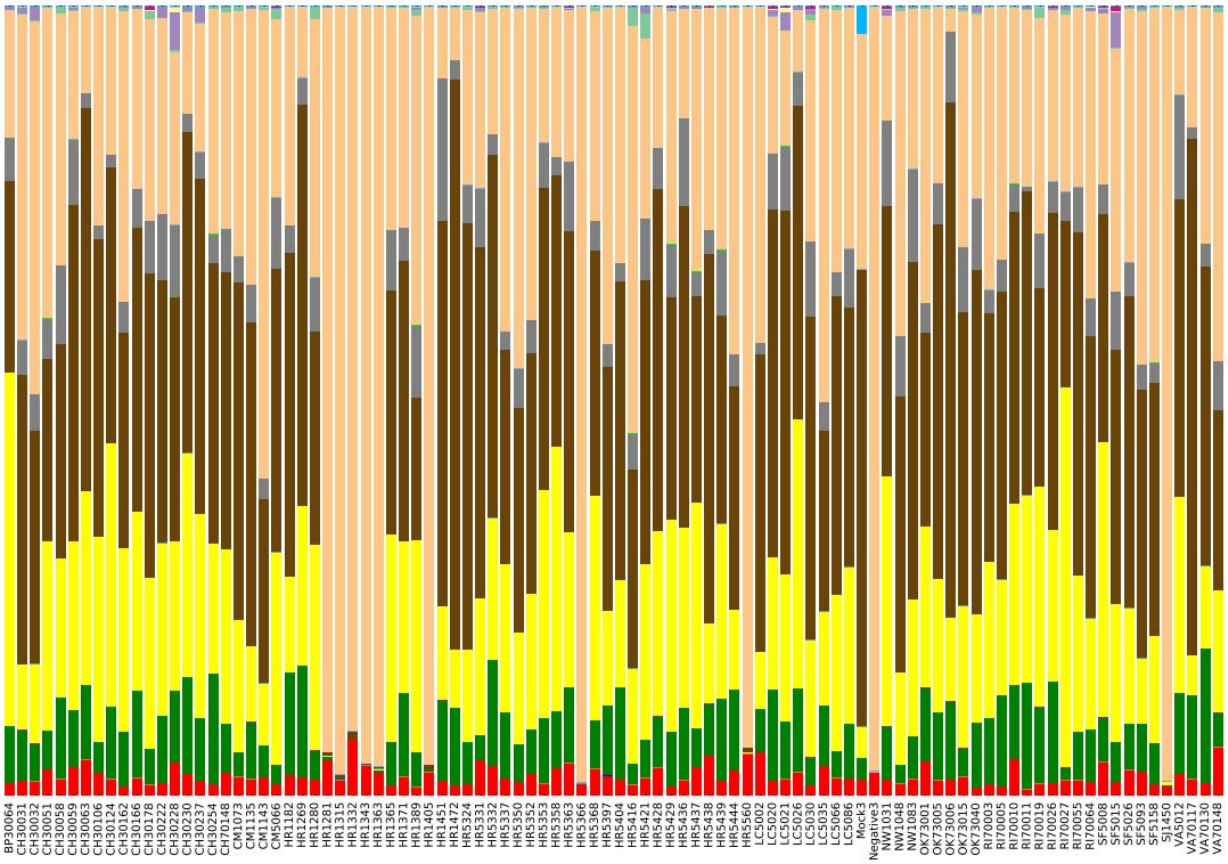


Figura 5.4. Número de individuos en la muestra

Legend	Taxonomy	Total	BP30064	CH30031	CH30032	CH30051	CH30058	CH30059	CH30063	CH30106	CH30124	CH30162	CH30166	CH30178	CH30222	CH30228	CH30230	CH30237	CH30254	CH70148	CM1073	CM1135	CM1143	CM5066	HR1182	HR1269	HR1280
		%	%	%	%	%	%	%	%	%	%	%	%	%	%	%	%	%	%	%	%	%	%	%	%	%	%
	Unassigned;Other	2.5%	1.5%	1.8%	1.8%	3.3%	2.1%	3.6%	4.5%	2.9%	2.0%	1.1%	2.2%	1.4%	1.6%	4.2%	2.8%	1.9%	1.4%	2.8%	2.4%	2.1%	2.3%	1.4%	2.6%	2.4%	2.0%
	k__Archaea;p__Euryarchaeota	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
	k__Bacteria;Other	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
	k__Bacteria;p__Actinobacteria	6.8%	7.2%	6.4%	4.8%	5.0%	10.3%	7.3%	9.5%	3.9%	9.2%	6.9%	11.1%	4.6%	8.4%	9.1%	12.2%	7.9%	14.0%	6.3%	3.1%	7.2%	4.0%	2.4%	12.9%	14.1%	3.7%
	k__Bacteria;p__Aquificae	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
	k__Bacteria;p__Bacteroidetes	17.5%	44.9%	8.3%	10.1%	23.9%	17.6%	21.3%	24.5%	26.0%	33.3%	23.3%	22.6%	21.6%	21.9%	18.9%	28.4%	26.0%	16.5%	22.0%	16.7%	9.7%	7.9%	27.0%	12.2%	20.3%	26.1%
	k__Bacteria;p__Chlorobi	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
	k__Bacteria;p__Chloroflexi	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
	k__Bacteria;p__FBP	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
	k__Bacteria;p__Firmicutes	33.0%	24.3%	36.6%	29.5%	23.1%	27.1%	42.6%	48.5%	37.7%	35.0%	27.2%	36.0%	38.6%	33.3%	30.9%	40.6%	42.3%	35.6%	35.1%	42.8%	40.9%	23.3%	35.9%	40.9%	50.9%	27.0%
	k__Bacteria;p__Fusobacteria	4.8%	5.5%	4.5%	4.6%	5.2%	10.1%	8.4%	1.9%	1.7%	1.5%	3.9%	4.9%	6.6%	8.4%	9.2%	2.5%	3.4%	3.7%	5.5%	3.2%	4.9%	2.6%	9.1%	3.6%	3.3%	6.8%
	k__Bacteria;p__GN02	0.0%	0.0%	0.0%	0.1%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.1%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.1%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.1%
	k__Bacteria;p__OP3	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
	k__Bacteria;p__Planctomycetes	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
	k__Bacteria;p__Proteobacteria	34.7%	16.1%	41.3%	47.3%	39.3%	31.9%	16.3%	11.0%	27.6%	18.8%	36.8%	22.8%	25.6%	24.8%	21.8%	12.9%	16.3%	28.5%	27.5%	31.2%	34.9%	59.4%	22.6%	27.7%	9.1%	32.8%
	k__Bacteria;p__SR1	0.2%	0.0%	0.0%	0.1%	0.1%	0.8%	0.1%	0.0%	0.0%	0.0%	0.4%	0.1%	1.0%	0.0%	0.4%	0.0%	0.1%	0.2%	0.5%	0.5%	0.1%	0.2%	1.3%	0.0%	0.0%	1.5%
	k__Bacteria;p__Spirochaetes	0.3%	0.4%	0.7%	1.7%	0.1%	0.1%	0.3%	0.0%	0.0%	0.0%	0.2%	0.3%	0.1%	1.4%	4.8%	0.6%	1.9%	0.1%	0.2%	0.1%	0.2%	0.1%	0.2%	0.0%	0.0%	0.0%
	k__Bacteria;p__Synergistetes	0.0%	0.0%	0.1%	0.1%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.7%	0.1%	0.1%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
	k__Bacteria;p__TM7	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
	k__Bacteria;p__Tenericutes	0.1%	0.1%	0.1%	0.0%	0.0%	0.0%	0.1%	0.0%	0.0%	0.0%	0.0%	0.0%	0.5%	0.0%	0.1%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.1%	0.2%	0.0%	0.0%	0.1%
	k__Bacteria;p__[Thermi]	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%

Figura 5.5. Porcentaje de cada individuo de la muestra

En la siguiente gráfica de la Figura 5.6 podemos ver los mismos individuos, pero esta vez clasificados por el sexo de las personas a las que pertenecen las muestras del experimento.

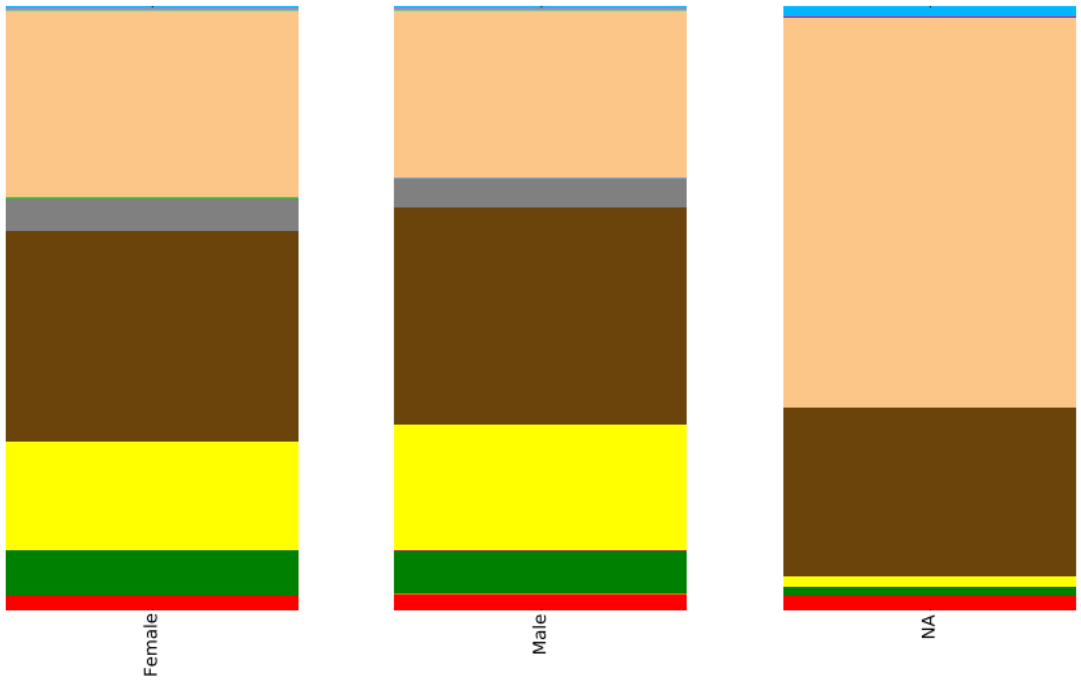


Figura 5.6. Clasificación de los individuos según los sexos de la muestra

Legend	Taxonomy	Total %	Female %	Male %	NA %
	Unassigned;Other	2.5%	2.4%	2.7%	2.5%
	k__Archaea;p__Euryarchaeota	0.0%	0.0%	0.0%	0.0%
	k__Bacteria;Other	0.0%	0.0%	0.0%	0.0%
	k__Bacteria;p__Actinobacteria	5.3%	7.5%	7.1%	1.3%
	k__Bacteria;p__Aquificae	0.0%	0.0%	0.0%	0.0%
	k__Bacteria;p__Bacteroidetes	13.6%	18.0%	21.0%	1.9%
	k__Bacteria;p__Chlorobi	0.0%	0.0%	0.0%	0.0%
	k__Bacteria;p__Chloroflexi	0.0%	0.0%	0.0%	0.0%
	k__Bacteria;p__FBP	0.0%	0.0%	0.0%	0.0%
	k__Bacteria;p__Firmicutes	32.9%	35.0%	35.9%	27.8%
	k__Bacteria;p__Fusobacteria	3.5%	5.4%	5.0%	0.0%
	k__Bacteria;p__GN02	0.0%	0.0%	0.0%	0.0%
	k__Bacteria;p__OP3	0.0%	0.0%	0.0%	0.0%
	k__Bacteria;p__Planctomycetes	0.0%	0.0%	0.0%	0.0%
	k__Bacteria;p__Proteobacteria	41.1%	30.9%	27.7%	64.8%
	k__Bacteria;p__SR1	0.2%	0.2%	0.3%	0.0%
	k__Bacteria;p__Spirochaetes	0.2%	0.4%	0.3%	0.0%
	k__Bacteria;p__Synergistetes	0.0%	0.0%	0.0%	0.0%
	k__Bacteria;p__TM7	0.0%	0.0%	0.0%	0.0%
	k__Bacteria;p__Tenericutes	0.0%	0.1%	0.1%	0.0%
	k__Bacteria;p__[Thermi]	0.6%	0.0%	0.0%	1.7%

Figura 5.7. Taxonomía de cada individuo según los sexos de la muestra

Y en esta última gráfica de la Figura 5.8 podemos ver la cantidad de individuos clasificados por la etnia de las personas que pertenecen las muestras del experimento.

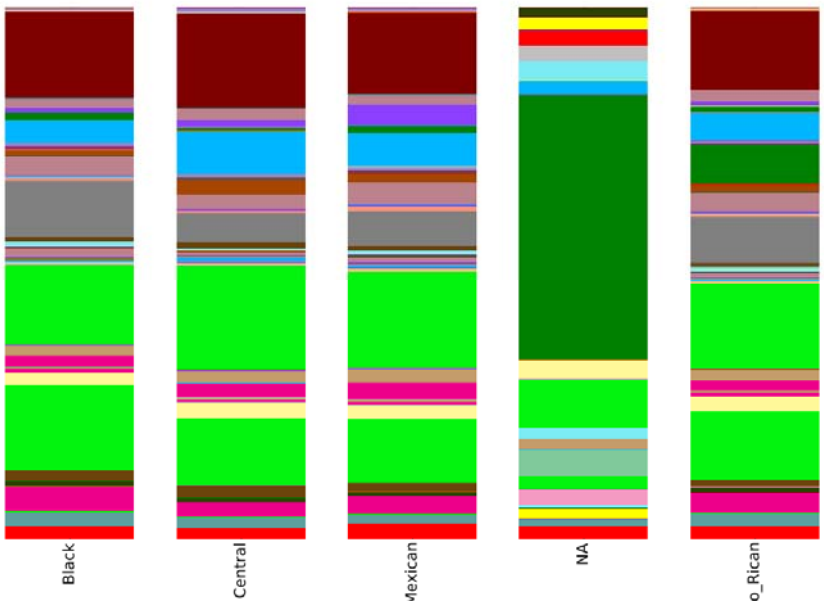


Figura 5.8. Número de individuos clasificados por las razas de la muestra

Legend	Taxonomy	Total %	Black %	Central %	Mexican %	NA %	Puerto_Rican %
	Unassigned;Other;Other;Other;Other	2.4%	2.4%	2.0%	2.8%	2.5%	2.5%
	k__Archaea;p__Euryarchaeota;c__Archaeoglobi;o__Archaeoglobales;f__Archaeoglobaceae	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
	k__Bacteria;Other;Other;Other;Other	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
	k__Bacteria;p__Actinobacteria;c__Acidimicrobiia;o__Acidimicrobiales;f__	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
	k__Bacteria;p__Actinobacteria;c__Actinobacteria;o__Actinomycetales;Other	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
	k__Bacteria;p__Actinobacteria;c__Actinobacteria;o__Actinomycetales;f__	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
	k__Bacteria;p__Actinobacteria;c__Actinobacteria;o__Actinomycetales;f__Actinomycetaceae	1.9%	2.5%	2.0%	1.7%	1.3%	2.2%
	k__Bacteria;p__Actinobacteria;c__Actinobacteria;o__Actinomycetales;f__Corynebacteriaceae	0.2%	0.3%	0.2%	0.3%	0.0%	0.1%
	k__Bacteria;p__Actinobacteria;c__Actinobacteria;o__Actinomycetales;f__Dermabacteraceae	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
	k__Bacteria;p__Actinobacteria;c__Actinobacteria;o__Actinomycetales;f__Geodermatophilaceae	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
	k__Bacteria;p__Actinobacteria;c__Actinobacteria;o__Actinomycetales;f__Microbacteriaceae	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
	k__Bacteria;p__Actinobacteria;c__Actinobacteria;o__Actinomycetales;f__Micrococcaceae	2.9%	4.6%	2.7%	3.4%	0.0%	3.9%
	k__Bacteria;p__Actinobacteria;c__Actinobacteria;o__Actinomycetales;f__Nocardioidaceae	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
	k__Bacteria;p__Actinobacteria;c__Actinobacteria;o__Actinomycetales;f__Propionibacteriaceae	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
	k__Bacteria;p__Actinobacteria;c__Actinobacteria;o__Bifidobacteriales;f__Bifidobacteriaceae	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
	k__Bacteria;p__Actinobacteria;c__Coriobacteriia;o__Coriobacteriales;f__Coriobacteriaceae	0.7%	0.9%	0.7%	0.7%	0.0%	1.0%
	k__Bacteria;p__Aquificae;c__Aquificae;o__Aquificales;f__Hydrogenothermaceae	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
	k__Bacteria;p__Bacteroidetes;Other;Other;Other	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
	k__Bacteria;p__Bacteroidetes;c__Bacteroidia;o__Bacteroidales;Other	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
	k__Bacteria;p__Bacteroidetes;c__Bacteroidia;o__Bacteroidales;f__	0.0%	0.0%	0.1%	0.1%	0.0%	0.0%
	k__Bacteria;p__Bacteroidetes;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae	0.4%	0.0%	0.0%	0.0%	1.9%	0.0%
	k__Bacteria;p__Bacteroidetes;c__Bacteroidia;o__Bacteroidales;f__Porphyromonadaceae	1.5%	1.9%	2.3%	1.6%	0.0%	1.4%
	k__Bacteria;p__Bacteroidetes;c__Bacteroidia;o__Bacteroidales;f__Prevotellaceae	10.7%	16.1%	12.6%	12.0%	0.0%	12.8%
	k__Bacteria;p__Bacteroidetes;c__Bacteroidia;o__Bacteroidales;f__Rikenellaceae	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
	k__Bacteria;p__Bacteroidetes;c__Bacteroidia;o__Bacteroidales;f__S24-7	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
	k__Bacteria;p__Bacteroidetes;c__Bacteroidia;o__Bacteroidales;f__[Odoribacteraceae]	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
	k__Bacteria;p__Bacteroidetes;c__Bacteroidia;o__Bacteroidales;f__[Paraprevotellaceae]	2.1%	2.3%	3.0%	2.5%	0.0%	2.8%
	k__Bacteria;p__Bacteroidetes;c__Flavobacteriia;o__Flavobacteriales;f__Cryomorphaceae	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
	k__Bacteria;p__Bacteroidetes;c__Flavobacteriia;o__Flavobacteriales;f__Flavobacteriaceae	0.5%	0.7%	0.5%	0.7%	0.0%	0.7%
	k__Bacteria;p__Bacteroidetes;c__Flavobacteriia;o__Flavobacteriales;f__[Weeksellaceae]	0.3%	0.3%	0.3%	0.5%	0.0%	0.4%
	k__Bacteria;p__Bacteroidetes;c__Sphingobacteriia;o__Sphingobacteriales;f__Sphingobacteriaceae	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
	k__Bacteria;p__Bacteroidetes;c__[Saprospirae];o__[Saprospirales];f__Chitinophagaceae	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
	k__Bacteria;p__Chlorobi;c__OPB56;o__;f__	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
	k__Bacteria;p__Chloroflexi;c__Anaerolineae;o__Anaerolineales;f__Anaerolinaceae	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
	k__Bacteria;p__Chloroflexi;c__Anaerolineae;o__Caldilineales;f__Caldilineaceae	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
	k__Bacteria;p__FBP;c__;o__;f__	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
	k__Bacteria;p__Firmicutes;Other;Other;Other	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
	k__Bacteria;p__Firmicutes;c__Bacilli;Other;Other	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
	k__Bacteria;p__Firmicutes;c__Bacilli;o__Bacillales;Other	0.0%	0.0%	0.0%	0.0%	0.2%	0.0%
	k__Bacteria;p__Firmicutes;c__Bacilli;o__Bacillales;f__	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
	k__Bacteria;p__Firmicutes;c__Bacilli;o__Bacillales;f__Alicyclobacillaceae	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
	k__Bacteria;p__Firmicutes;c__Bacilli;o__Bacillales;f__Bacillaceae	0.7%	0.0%	0.0%	0.0%	3.4%	0.0%
	k__Bacteria;p__Firmicutes;c__Bacilli;o__Bacillales;f__Listeriaceae	0.5%	0.0%	0.0%	0.0%	2.6%	0.0%
	k__Bacteria;p__Firmicutes;c__Bacilli;o__Bacillales;f__Paenibacillaceae	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
	k__Bacteria;p__Firmicutes;c__Bacilli;o__Bacillales;f__Planococcaceae	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
	k__Bacteria;p__Firmicutes;c__Bacilli;o__Bacillales;f__Staphylococcaceae	1.0%	0.0%	0.0%	0.0%	5.0%	0.0%
	k__Bacteria;p__Firmicutes;c__Bacilli;o__Gemellales;Other	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
	k__Bacteria;p__Firmicutes;c__Bacilli;o__Gemellales;f__	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
	k__Bacteria;p__Firmicutes;c__Bacilli;o__Gemellales;f__Gemellaceae	2.0%	2.0%	2.7%	3.1%	0.0%	1.9%
	k__Bacteria;p__Firmicutes;c__Bacilli;o__Lactobacillales;Other	2.1%	1.9%	2.3%	2.4%	1.9%	2.0%
	k__Bacteria;p__Firmicutes;c__Bacilli;o__Lactobacillales;f__	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
	k__Bacteria;p__Firmicutes;c__Bacilli;o__Lactobacillales;f__Aerococcaceae	0.1%	0.1%	0.1%	0.2%	0.0%	0.1%

Figura 5.9. Taxonomía de los individuos según las razas de la muestra

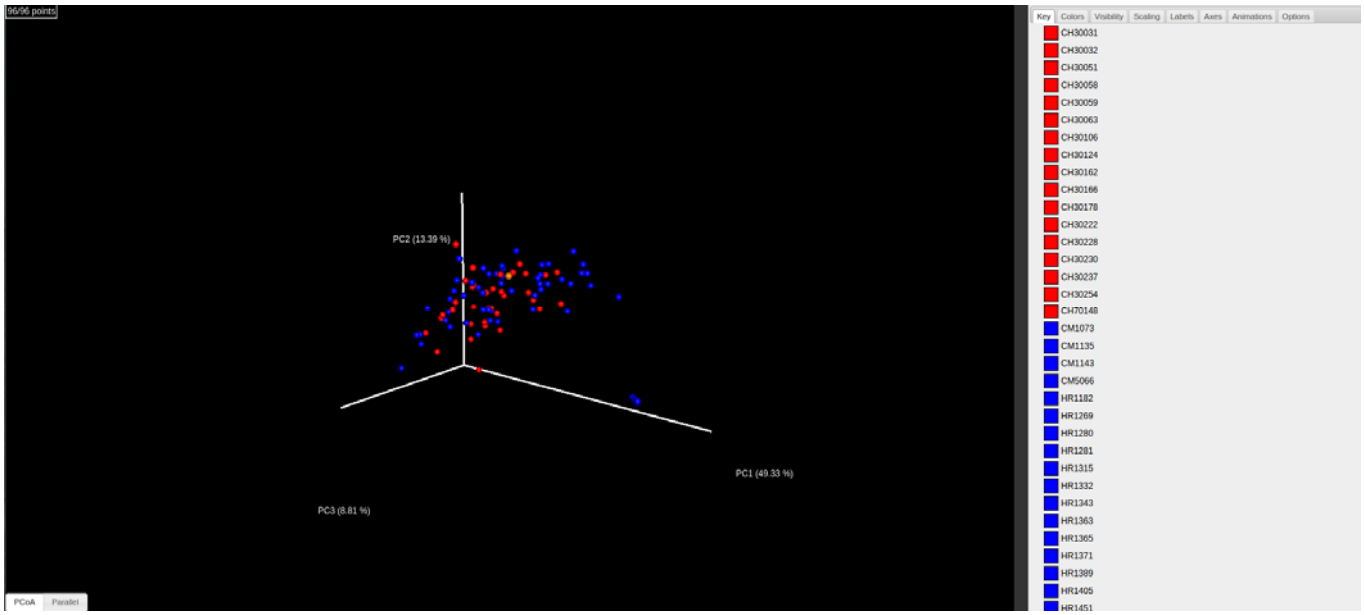


Figura 5.10. Visualización tridimensional de los resultados

También podemos observar una representación tridimensional sobre la distribución de los resultados de la muestra (ver Figura 5.10). Esta gráfica es de una herramienta externa, **EMPeror** [18], que está integrada en QIIME, y permite visualizaciones tridimensionales interactivas sobre los resultados obtenidos del análisis. Podemos filtrar por las mismas categorías que tenemos en las gráficas anteriores.

5.2.2 Resultados de PICRUST

En los resultados de PICRUST, lo que esperamos observar, son los procesos llevados a cabo por los individuos de una comunidad microbiana. En este caso tenemos dichas funciones representadas en dos gráficas distintas, una de barras (Figura 5.11) y otra de área (Figura 5.12), que representan de manera gráfica, el porcentaje de la función X dentro de una muestra determinada (Figura 5.13).

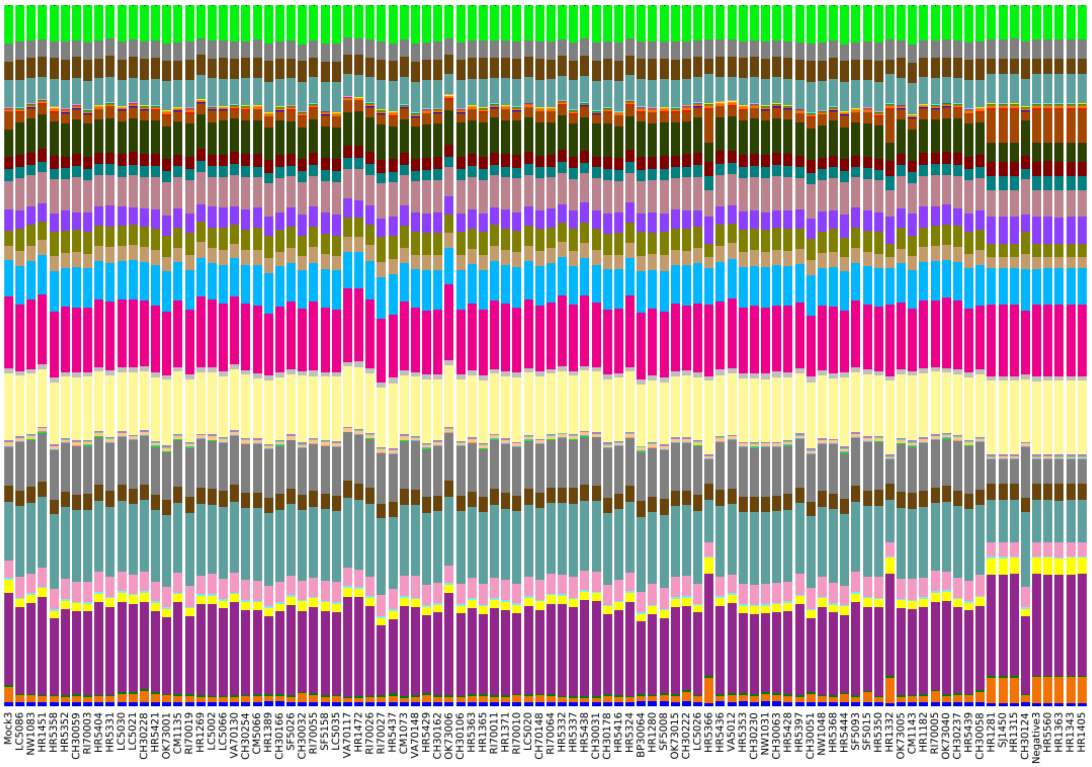


Figura 5.11. Resultados de PICRUSt en diagrama de barras

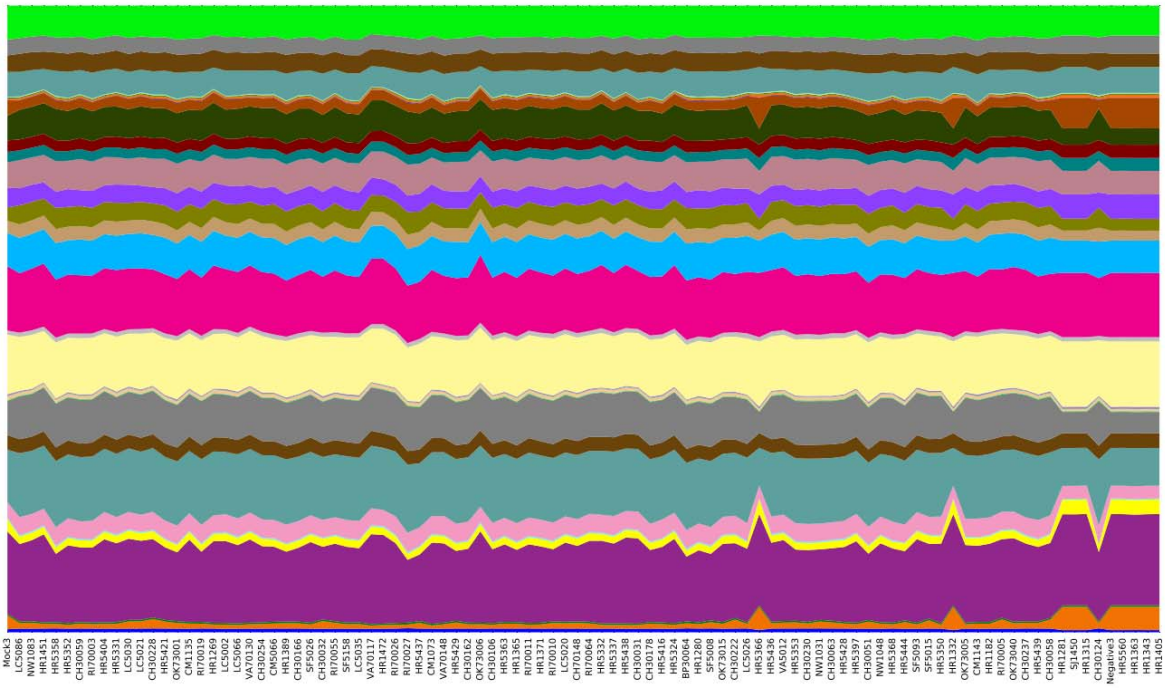


Figura 5.12. Resultados de PICRUSt en formato de áreas

Legend	Taxonomy	Total	Mock3	LC5086	NW1083	HR1451	HR5358	HR5352	CH30059	RI70003	HR5404	HR5331	LC5030	LC5021	CH30228	HR5421	OK73001	CM1135	RI70019	HR1269
		%	%	%	%	%	%	%	%	%	%	%	%	%	%	%	%	%	%	%
	Cellular Processes;Cell Communication	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
	Cellular Processes;Cell Growth and Death	0.6%	0.5%	0.6%	0.6%	0.6%	0.6%	0.6%	0.6%	0.6%	0.6%	0.6%	0.6%	0.6%	0.6%	0.6%	0.6%	0.6%	0.6%	0.6%
	Cellular Processes;Cell Motility	1.1%	2.2%	0.9%	0.9%	0.8%	0.7%	0.6%	0.8%	0.7%	0.8%	0.8%	1.2%	1.2%	1.5%	1.1%	0.9%	0.8%	0.8%	0.7%
	Cellular Processes;Transport and Catabolism	0.3%	0.3%	0.3%	0.3%	0.2%	0.3%	0.3%	0.2%	0.3%	0.3%	0.3%	0.2%	0.2%	0.3%	0.2%	0.3%	0.2%	0.3%	0.3%
	Environmental Information Processing;Membrane Transport	12.4%	13.1%	12.3%	12.9%	14.1%	10.9%	12.4%	11.9%	12.0%	13.3%	12.6%	12.9%	12.6%	12.5%	11.6%	11.0%	13.1%	11.2%	13.0%
	Environmental Information Processing;Signal Transduction	1.3%	1.9%	1.2%	1.2%	1.2%	1.0%	1.1%	1.1%	1.1%	1.2%	1.2%	1.3%	1.2%	1.3%	1.2%	1.2%	1.2%	1.1%	1.1%
	Environmental Information Processing;Signaling Molecules and Interaction	0.2%	0.3%	0.2%	0.2%	0.2%	0.2%	0.2%	0.2%	0.2%	0.2%	0.2%	0.2%	0.2%	0.2%	0.2%	0.2%	0.2%	0.2%	0.3%
	Genetic Information Processing;Folding, Sorting and Degradation	2.8%	2.4%	2.9%	2.8%	2.7%	3.0%	3.0%	2.8%	3.0%	2.9%	2.8%	2.8%	2.7%	2.9%	2.9%	2.9%	2.8%	3.0%	2.7%
	Genetic Information Processing;Replication and Repair	9.8%	8.4%	10.2%	10.1%	10.1%	10.6%	10.2%	10.3%	10.3%	10.1%	10.1%	10.1%	10.1%	10.2%	10.2%	10.2%	9.9%	10.5%	10.2%
	Genetic Information Processing;Transcription	2.2%	2.4%	2.2%	2.3%	2.3%	2.2%	2.2%	2.1%	2.2%	2.3%	2.3%	2.2%	2.2%	2.2%	2.2%	2.2%	2.3%	2.2%	2.3%
	Genetic Information Processing;Translation	6.5%	5.4%	6.8%	6.8%	7.0%	7.0%	6.9%	7.0%	6.9%	6.9%	6.8%	6.8%	6.9%	6.9%	6.9%	6.8%	6.8%	7.0%	6.9%
	Human Diseases;Cancers	0.1%	0.1%	0.1%	0.1%	0.1%	0.1%	0.1%	0.1%	0.1%	0.1%	0.1%	0.1%	0.1%	0.1%	0.1%	0.1%	0.1%	0.1%	0.1%
	Human Diseases;Cardiovascular Diseases	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
	Human Diseases;Immune System Diseases	0.1%	0.1%	0.1%	0.1%	0.1%	0.1%	0.1%	0.1%	0.1%	0.1%	0.1%	0.1%	0.1%	0.1%	0.1%	0.1%	0.1%	0.1%	0.1%
	Human Diseases;Infectious Diseases	0.4%	0.5%	0.4%	0.4%	0.4%	0.4%	0.4%	0.4%	0.4%	0.4%	0.4%	0.4%	0.4%	0.4%	0.4%	0.4%	0.4%	0.4%	0.4%
	Human Diseases;Metabolic Diseases	0.1%	0.1%	0.1%	0.1%	0.1%	0.1%	0.1%	0.1%	0.1%	0.1%	0.1%	0.1%	0.1%	0.1%	0.1%	0.1%	0.1%	0.1%	0.1%
	Human Diseases;Neurodegenerative Diseases	0.2%	0.3%	0.2%	0.2%	0.1%	0.1%	0.2%	0.2%	0.2%	0.2%	0.2%	0.2%	0.2%	0.2%	0.2%	0.3%	0.2%	0.2%	0.1%
	Metabolism;Amino Acid Metabolism	9.0%	9.5%	8.6%	8.6%	8.2%	8.9%	8.7%	9.0%	9.0%	8.5%	8.8%	8.5%	8.9%	8.5%	9.1%	9.4%	8.8%	8.9%	8.7%
	Metabolism;Biosynthesis of Other Secondary Metabolites	0.7%	0.7%	0.7%	0.7%	0.7%	0.8%	0.7%	0.7%	0.7%	0.7%	0.7%	0.7%	0.7%	0.7%	0.7%	0.7%	0.7%	0.7%	0.8%
	Metabolism;Carbohydrate Metabolism	9.6%	10.2%	9.5%	9.8%	10.1%	9.2%	9.4%	9.3%	9.2%	9.7%	9.9%	9.7%	9.7%	9.4%	9.4%	9.1%	9.6%	9.3%	10.1%
	Metabolism;Energy Metabolism	5.5%	5.3%	5.6%	5.5%	5.5%	5.8%	5.5%	5.7%	5.6%	5.4%	5.5%	5.5%	5.6%	5.5%	5.8%	5.7%	5.4%	5.7%	5.5%
	Metabolism;Enzyme Families	2.1%	1.9%	2.2%	2.2%	2.1%	2.3%	2.2%	2.2%	2.1%	2.2%	2.2%	2.1%	2.2%	2.2%	2.1%	2.1%	2.1%	2.2%	2.2%
	Metabolism;Glycan Biosynthesis and Metabolism	2.9%	2.2%	3.0%	2.9%	2.7%	3.4%	3.2%	3.2%	3.1%	3.0%	3.0%	2.8%	2.9%	2.9%	3.0%	3.0%	2.8%	3.1%	2.8%
	Metabolism;Lipid Metabolism	2.8%	3.1%	2.7%	2.7%	2.6%	2.6%	2.8%	2.5%	2.7%	2.7%	2.9%	2.9%	2.6%	2.6%	2.7%	3.0%	2.8%	2.7%	2.6%
	Metabolism;Metabolism of Cofactors and Vitamins	4.5%	4.1%	4.5%	4.4%	4.4%	4.9%	4.5%	5.0%	4.7%	4.5%	4.4%	4.1%	4.5%	4.5%	4.7%	4.7%	4.4%	4.8%	4.5%
	Metabolism;Metabolism of Other Amino Acids	1.6%	1.7%	1.5%	1.5%	1.6%	1.6%	1.5%	1.6%	1.6%	1.5%	1.6%	1.5%	1.6%	1.5%	1.6%	1.6%	1.6%	1.5%	1.6%
	Metabolism;Metabolism of Terpenoids and Polyketides	1.8%	1.8%	1.7%	1.7%	1.8%	1.8%	1.7%	1.7%	1.7%	1.7%	1.7%	1.7%	1.7%	1.7%	1.8%	1.8%	1.7%	1.8%	1.8%
	Metabolism;Nucleotide Metabolism	4.6%	3.9%	4.7%	4.8%	4.9%	5.0%	4.8%	4.9%	4.8%	4.8%	4.8%	4.6%	4.8%	4.8%	4.8%	4.7%	4.7%	4.8%	5.0%
	Metabolism;Xenobiotics Biodegradation and Metabolism	1.8%	2.5%	1.5%	1.6%	1.6%	1.4%	1.5%	1.4%	1.4%	1.4%	1.7%	1.6%	1.6%	1.5%	1.5%	1.6%	1.6%	1.5%	1.6%
	Organismal Systems;Circulatory System	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
	Organismal Systems;Digestive System	0.1%	0.0%	0.1%	0.1%	0.1%	0.1%	0.1%	0.1%	0.1%	0.1%	0.1%	0.1%	0.1%	0.1%	0.1%	0.1%	0.1%	0.1%	0.1%
	Organismal Systems;Endocrine System	0.3%	0.2%	0.3%	0.3%	0.3%	0.3%	0.3%	0.3%	0.3%	0.3%	0.3%	0.3%	0.3%	0.3%	0.3%	0.3%	0.3%	0.3%	0.3%
	Organismal Systems;Environmental Adaptation	0.1%	0.1%	0.1%	0.1%	0.1%	0.1%	0.1%	0.1%	0.1%	0.1%	0.1%	0.1%	0.1%	0.1%	0.1%	0.1%	0.1%	0.1%	0.1%
	Organismal Systems;Excretory System	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%

Figura 5.13. Taxonomía de las funciones según cada uno de los individuos

5.2.3 Resultados de FastQC

Por último, tenemos los resultados de FastQC, en los que podemos observar, primero, una representación gráfica de las muestras antes y después del filtrado, y segundo, la evolución de las lecturas (ver Figura 5.14 y Figura 5.15).

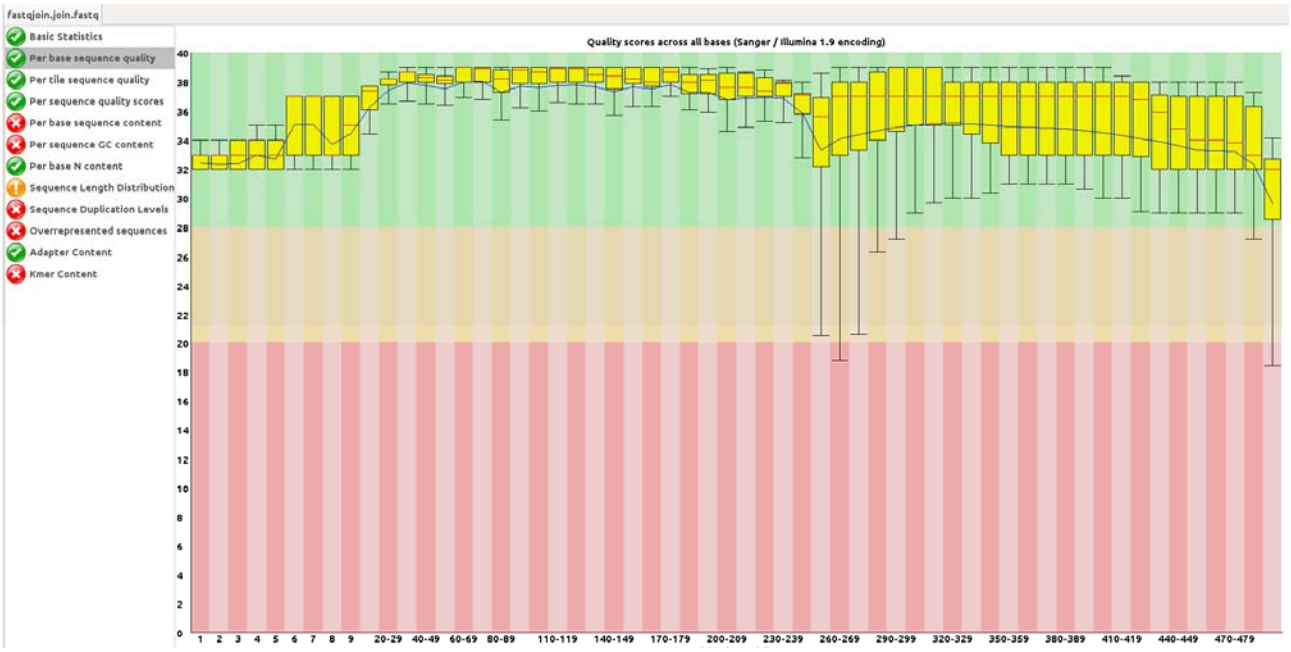


Figura 5.14. Representación de las muestras antes del filtrado

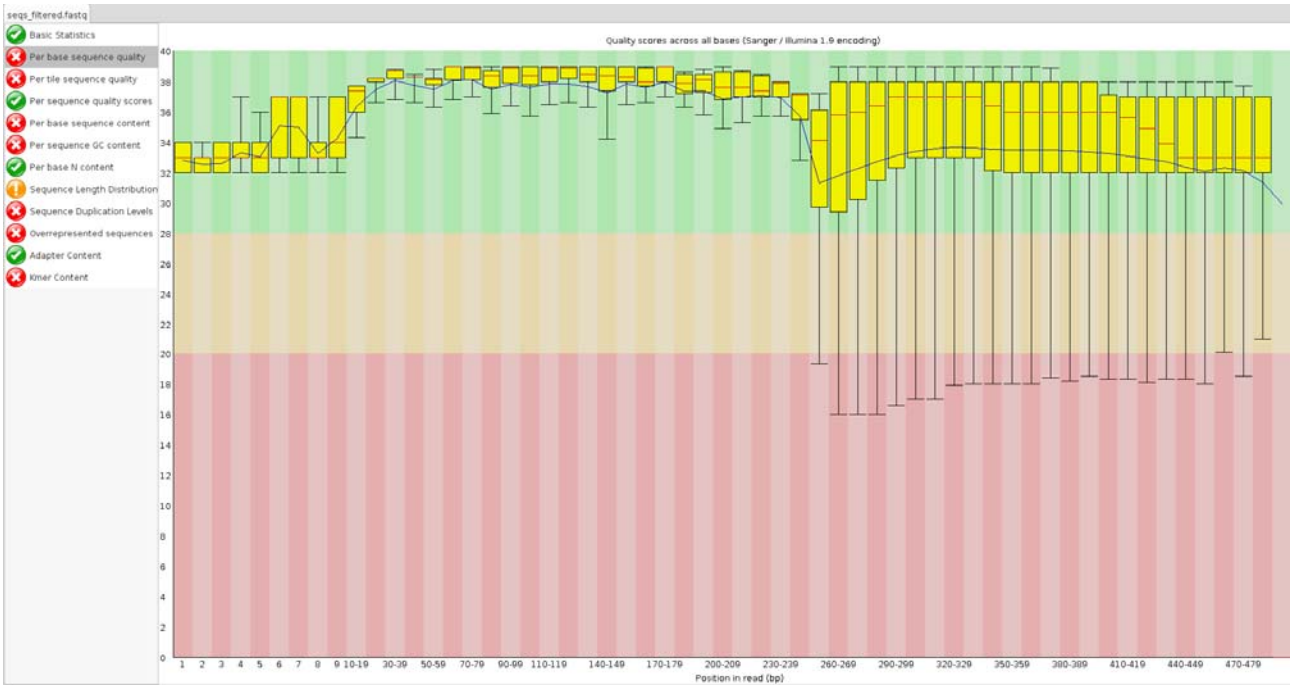


Figura 5.15. Representación de las muestras después del filtrado

Capítulo 6.
Conclusiones y líneas futuras

En el presente trabajo se describe el desarrollo de QiimeApp, una plataforma web para llevar a cabo análisis metagenómicos de secuencias de ADN microbiano, partiendo de datos secuenciados por equipos de Illumina o de Ion Torrent. Durante el desarrollo de la aplicación, nos hemos basado en lecturas procedentes de muestras de saliva, para obtener un funcionamiento óptimo de la herramienta y conseguir los resultados esperados.

Para llevar a cabo el procesamiento y análisis de los datos extraídos, nos decantamos por la utilización de la herramienta *open-source* QIIME. Durante la exploración acerca de su funcionamiento, nos encontramos con multitud de problemas, la mayoría de ellos a la hora de realizar una correcta instalación de la herramienta en los terminales de trabajo. Al final, aunque se consiguió instalar satisfactoriamente en uno de los ordenadores pertenecientes al personal investigador del Hospital Universitario Nuestra Señora de La Candelaria, se optó, por utilizar una máquina virtual completamente configurada con QIIME y perfectamente funcional, que es proporcionada por los propios responsables y autores de QIIME a través de su página web.

A pesar de la escueta documentación oficial, una vez que pudo ser entendido el funcionamiento de QIIME, así como de los scripts que la componen, el siguiente paso fue hacer una selección de los scripts que eran necesarios para el proyecto. A continuación, a partir de esta lista, se pasó a desarrollar un pipeline en Bash, que sería el encargado de procesar y analizar los datos que el usuario le pasara.

Con el pipeline en Bash funcionando, se decidió abstraer de su utilización al usuario, y para ello, se crearía una aplicación web que enmascararía al pipeline, que funcionaría a través de dicha capa web, y que fuera la responsable de interactuar directamente con el usuario. La aplicación web se implementó usando el *framework* de desarrollo web Django, así como también

HTML, CSS y Javascript, siguiendo los estándares establecidos para un correcto desarrollo de la aplicación.

También se decidió complementar la información aportada en los resultados de QIIME, por otras dos herramientas bioinformáticas: PICRUST y FastQC. La utilidad de PICRUST residía en la posibilidad de realizar Metagenómica funcional, es decir, poder predecir las funciones celulares llevadas a cabo por los organismos que componen una comunidad microbiana. FastQC, por otro lado, es una herramienta para la visualización de las distribuciones de calidades y longitudes asociadas a las lecturas de un experimento. Con FastQC podemos explorar fácilmente las diferencias existentes en las lecturas antes y después de ser filtradas y así complementar con información extra los resultados obtenidos de los análisis principales llevados a cabo por QIIME.

La herramienta, QiimeApp, ha sido capaz pues, de resolver el problema inicialmente planteado, el de la falta de una herramienta apropiada para poder realizar análisis metagenómicos, en investigaciones llevadas a cabo por el personal del Hospital Universitario Nuestra Señora de la Candelaria y de la Unidad de Genómica Aplicada de la Universidad de La Laguna. La herramienta proporciona una forma cómoda y simple de realizar distintos análisis metagenómicos de manera completamente autónoma y transparente para el usuario.

Como posibles mejoras y ampliaciones futuras para la herramienta, se considera ampliar las opciones de configuración disponibles para poder aceptar distintos tipos proyectos. Otra posible ampliación a destacar, sería la de añadir más herramientas al pipeline, de manera que enriquezcan notablemente los resultados finales obtenidos, sobre todo de representación gráfica. Cuanta más y mejor información se le pueda dar en el resultado final al usuario, mejores serán las conclusiones que pueda sacar en las investigaciones que lleve a cabo.

Otra opción de mejora posible, es que dado el consumo de recursos que se requiere para su utilización, una de las líneas futuras más interesantes sería la de, en lugar de como hasta ahora, donde el usuario ha de instalarse todo lo necesario para poder trabajar con la herramienta, poder integrarla bien en servidores locales, del centro de investigación para el uso de su personal y poder trabajar así desde el navegador web de sus propios ordenadores o

terminales de trabajo, o bien integrarla en un servicio en la nube, para su acceso y uso por parte de cualquier persona desde cualquier parte del mundo.

Capítulo 7. Presupuesto

Para el cálculo del presupuesto total, hemos tenido en cuenta los siguientes elementos descritos en la Tabla 7.1.

Tipo	Cantidad	Descripción
Licencia del software utilizado para el proyecto	0 €	Licencias *open-source*
Otras Licencias de software	126€	Licencia Microsoft Office 2016
Adquisición de hardware	143€	Mejorar las prestaciones de los terminales de trabajo
Servidor	3000€	Servidor local para poner en funcionamiento la plataforma
Personal de mantenimiento	15€/hora	Personal encargado del mantenimiento de los servidores

Tabla 7.1. Presupuesto de la aplicación

Capítulo 8.
Summary and Conclusions

In this work we describe the development of QiimeApp, a web platform to perform metagenomic analysis of microbial DNA sequences, using data sequenced by Illumina equipment or Ion Torrent. During the application development, we have taken the sequencing reads from saliva samples, for the optimum operation of the tool and to achieve the expected results.

To carry out the processing and analysis of the extracted data, we decided to use the open-source tool called QIIME. During the exploration of its operation, we encountered many problems, most of them when we tried to install it properly on the workstations. Although we managed to install the app successfully on one of the computers belonging to the research staff of the *Hospital Universitario Nuestra Señora de La Candelaria*, at the end, we decided to use a virtual machine, which is fully configured with QIIME and completely functional. The virtual machines are provided by the authors and developers of QIIME, through their website.

Despite the fact of the poor official documentation, once we could understand how QIIME works and its scripts, the next step was to make a list of all the scripts which were necessary for the project. Then, with the list of this scripts, it was decided to develop a pipeline in Bash, which it would be responsible of managing and analyzing all the data from the users.

Once we had developed the pipeline, the next step was to create a web application, which it would hide the pipeline to the user, so he/she would only need to interact with the web layer and not with the pipeline itself. The web application was developed using the Django web framework, as well as HTML, CSS and Javascript, following the official standards for a proper development of the application.

Furthermore, we decided to complement the information, which is provided by QIIME in the final results, with two more bioinformatic tools: PICRUSt and FastQC. PICRUSt allow us to perform functional metagenomics, which meant that we were are able to predict cellular functions carried out by the

organism living inside of a microbial community. FastQC is a tool for displaying quality distributions and the reads lengths from a specific experiment. With FastQC we could see the differences between the reads before and after being filtered in order to complement the results of QIIME.

To summarize, QiimeApp has been able to solve the initial problem we faced, the lack of a suitable tool to perform metagenomic analysis on the research conducted by the staff of the *Hospital Universitario Nuestra Señora de la Candelaria* and the Applied Genomics Group at the University of La Laguna. The tool provides a convenient and easy way to perform different metagenomic analysis, which are fully autonomous and transparent for the user.

As future improvements and extensions for the tool, we consider to expand the configuration options available in order to accept different types projects. Another possible extension, would be to add more bioinformatics tools to the pipeline, particularly of additional plotting tools, to get better information in the final results. The better information we provide to the user, the better would be the final conclusions of his current research.

Another possible improvement, considering the resource consumption required for a proper use of the tool, it is to run the application on servers, instead of letting the user to install all the components of the application in their own computers and deal with all the problems. It is one of the most interesting improvements for the application. There are two possibilities: the first is to install it on local servers in the research center in order to be used by the local researchers using a web browser from their computers; whereas the second one is to make use of the cloud computing services available, like Amazon o Google, and let everyone access the application from different places and countries.

Bibliografía

[1] Metagenómica

en.wikipedia.org/wiki/Metagenomics,

Torsten Thomas, Jack Gilbert and Folker Meyer (2012)

Metagenomics – a guide from sampling to data analysis

www.ncbi.nlm.nih.gov/pmc/articles/PMC3351745/.

[2] Next Generation Sequencer
www.illumina.com/content/dam/illumina-marketing/documents/products/illumina_sequencing_introduction.pdf

[3] PGM, Ion Torrent sequencer
www.thermofisher.com/order/catalog/product/4462921

[4] Miseq, Illumnia sequencer http://www.illumina.com/systems/miseq.html

[5] 16S rRNA. es.wikipedia.org/wiki/ARN_ribosomal_16S,
www.ncbi.nlm.nih.gov/pmc/articles/PMC523561/

[6] IonGap iongap.hpc.iter.es/iongap

[7] EBI, metagenomics webserver www.ebi.ac.uk/metagenomics/

[8] SILVA NGS www.arb-silva.de

[9] MG-RAST metagenomics.anl.gov

[10] QIIME qiime.org

[11] MOTHUR www.mothur.org

[12] OTUs en.wikipedia.org/wiki/Operational_taxonomic_unit
www.metagenomics.wiki/pdf/definition/operational-taxonomic-unit-otu

[13] UCLUST drive5.com/usearch/manual/uclust_algo.html

[14] PICRUSt picrust.github.io/picrust

[15] FastQC www.bioinformatics.babraham.ac.uk/projects/fastqc

[16] Django www.djangoproject.com

[17] *An evaluation of the accuracy and speed of metagenome analysis tools*

Stinus Lindgreen, Karen L. Adair, Paul P. Gardner (2016)

www.nature.com/articles/srep19233

[18] EMPeror github.com/biocore/emperor