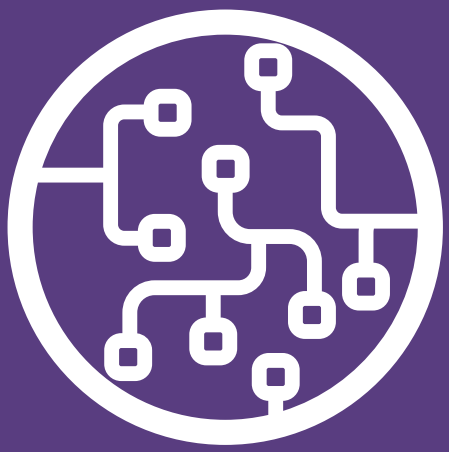
UMAG
Universidad de Magallanes

Jacqui Tesis



Jacqueline Aldridge Águila

UMAG
Universidad de Magallanes

Jacqui Tesis

Jacqueline Aldridge Águila

Dirigida por
Dr. Roberto Uribe-Paredes

Tesis para optar al grado de

Magíster en Bioinformática

Departamento de Ingeniería en Computación
Facultad de Ingeniería
Universidad de Magallanes

Julio, 2024

Índice general

P

Declaración de Autenticidad

P

UMAG
Universidad de Magallanes

Declaro que la presente tesis y el trabajo presentado en ella son de mi propia autoría. Basado en mi comprensión y conocimiento, puedo afirmar que este trabajo es original y, en aquellos casos donde se han desarrollado ideas en colaboración con otras personas, se han realizado las citas y referencias apropiadas para reconocer dichas contribuciones. Finalmente, confirmo que este trabajo no ha sido presentado para ningún otro grado o calificación académica.

Título	Jacqui Tesis
Autor	Jacqueline Aldridge Águila
Grado	Magíster en Bioinformática
Facultad	Facultad de Ingeniería

Fecha Julio, 2024

Agradecimientos

Quiero expresar mi más profundo agradecimiento a mi asesor por su invaluable orientación, paciencia y apoyo a lo largo de este proceso de investigación. Mi gratitud se extiende a los miembros de mi comité, por sus perspicaces comentarios y sugerencias. Agradezco también a mi familia y amigos por su amor incondicional y aliento en los momentos más desafiantes. Este trabajo no habría sido posible sin el apoyo y la motivación de todas estas personas.

También muchas gracias a LaTeX [1] por su gran ayuda en la redacción de este documento.

Disclaimer: Contenido generado con ChatGPT.

1

Introducción

2

Marco teórico y estado del arte

2.1 | biologia

Microbiota

La microbiota es el conjunto de microorganismos (bacterias, virus, arqueas, u hongos) que habitan en un ambiente, ya sea en organismos multicelulares como humanos [2], animales [3] o plantas [4], o en ambientes naturales como el océano [5] o el suelo [6]. Estos organismos presentes en la microbiota se encuentran en un estado de simbiosis junto con el host, contribuyendo en funciones vitales como la homeostasis, regulación del sistema inmune, digestión de alimentos, producción de vitaminas, protección ante enfermedades y agentes patógenos [7–10]. Sin embargo, una disbiosis o una baja diversidad en la microbiota puede llevar a una desregulación del organismo, incluyendo diversos tipos de enfermedades, fallas en el sistema inmune, falta de vitaminas, trastornos como depresión, estrés, e incluso diferentes tipos de cáncer en el caso del ser humano [9, 10].

La composición de la microbiota va cambiando dependiendo del área que se está colonizando, pudiéndose ecntrar diferentes microorganismos en las cavidades orales, zonas intestinales, genitales, cutáneas o tracto respiratorio [11].

Se estima que en el ser humano habitan más de 10 billones de microorganismos [12], es decir, poseemos cerca de 350 billones de celulas microbioanas [8, 13] , siendo este número al menos 10 veces mayor que el número de células humanas que poseemos.

En la naturaleza los microorganismos cumplen un rol fundamental en los ciclos bioquímicos del nitrógeno, carbono y fósforo [14, 15], como también en los procesos de desnitrificación, nitrificación y mineralización [14, 15]. Dependiendo del tipo de ambiente, los microorganismos también varian, en el caso del suelo por ejemplo, cambian dependiendo del tipo de suelo en el que están (agrícolas, forestales, humedales, pastos o suelos desérticos [16]) y de las características de éste como la temperatura, hidaratación, profundidad, cantidad de carbono [17]. En el caso de las plantas, se ha demostrado, que la microbiota presente ayuda a la adquisición de nutrientes [18], al crecimiento,

salud de las plantas y resistencia a enfermedaddes [19–22].

La microbiota humana se puede ver afectada por diferentes factores, como los hábitos alimenticios, estilo de vida, uso de antibióticos, la edad, estrés, entre otros [9]. La interacción con el medio ambiente influye notoriamente también, habiendo estudios que identifican cambios en la microbiota intestinal y cutánea en niños que interactuan con la naturaleza, plantas o suelo, identificando también un aumento de vías inmunoreguladoras en comunidades microbianas cercanas a la naturaleza [23]. También se han identificado cambios en la microbiota de recién nacidos, infantes y adultos que viven con animales [24–26].

Conocer la diversidad microbiana asociada a organismos multicelulares permite conocer microorganismos patógenos que causan enfermedades infecciosas,lo cual ayudaría al diagnóstico y permitiría tomar acciones oportunas [27, 28].

ARN Ribosomal 16S

El gen 16S rRNA es el marcador molecular más utilizado para la identificación de bacterias y comunidades microbianas [29, 30]. Posee caracteristicas únicas, como su presencia en todas las bacterias, su alto grado de conservación (debido a que su función no cambia a través del tiempo) y su tamaño, el cuál permite ser lo suficientemente largo y preciso para la asignación taxónomica, y abordable para análisis bioinformáticos [29, 31].

El gen 16S rRNA esta compuesto de aproximadamente 1542 pares de bases divididas en 10 regiones conservadas y 9 regiones hipervariables [32].

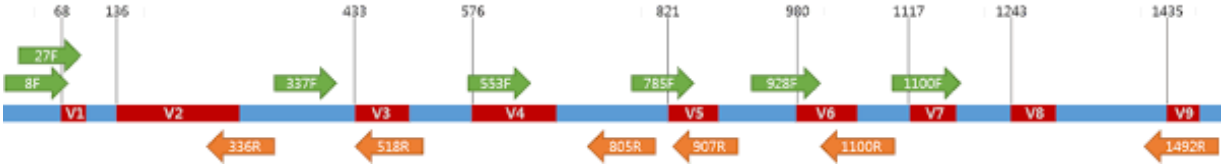


Figura 2.1. Estructura de las regiones variables e hipervariables del gen 16S rRNA

Figura 2.2. Estructura de las regiones variables e hipervariables del gen 16S rRNA

hacer denuevo la figura

Las regiones hipervariables permiten llevar a cabo la caracterización de los microorganismos. Diversos estudios se han llevado a cabo para determinar los efectos de la selección de la región a utilizar para la identificación, llegandose a determinar que la región hipervariable ha utilizar influye en los resultados de la comunidad y en la diversidad de organismos que se caracteriza [33–36], y también que ciertas regiones hipervariables permiten identificar mejor ciertos grupos taxonómicos [**buscar**].

Con el desarrollo de las tecnologías de secuenciación masiva, su bajo costo y alto througput, la forma de caracterizar bacterias se ha vuelto más estándar y abordable al día de hoy [37, 38]. El estudio de comunidades microbianas mediante el gen 16S rRNA se ha vuelto una herramienta poderosa tanto en ambientes clínicos, como ambientales, incluso llegandose a secuenciar en el espacio, permite obtener información de la diversidad de una muestra de manera mucho más rápida y económica que los métodos tradicionales [**buscar**].

Secuenciación de ADN

que es la secuenciaciónd de adn Mediante la secuencación de ADN se puede caracterizar genomas completos,

Con la aparición de las tecnologías de secuenciación de segunda y tercera generación [29, 39], la secuenciación del gen 16S rRNA se convirtió en es una técnica masiva hoy en día para la caracterización de comunidades microbioanas e identificación tanto de patogenos o aislamiento de bacterias clínicas [31].

Estos métodos requieren la amplificación y secuenciación del gen 16S rRNA y el uso de herramientas biofinformáticas para la identificación y comparación con bases de datos.

Con la secuenciación del gen 16S rRNA se obtiene un conjunto de lecturas de ADN, donde cada lectura pertenece a una bacteria presente en la muestra. Estas lecturas se procesan y se comparán con bases de datos existentes para poder realizar la asignación taxonómica y poder identificar la bacteria. Finalmente lo que se obtiene es un perfil de toda la comunidad bacteriana de la muestra, todas las bacterias presentes que las herramientas bioinformáticas pudieron detectar, junto con su abundancia relativa.

Existen diferentes tecnologías de secuenciación cada una de ellas con diferentes largos de las lecturas a secuenciar,diferente porcentaje de error, costo y througput. Cada una de ellas tiene sus ventajas y desventajas, y los análisis bioinformáticos a realizar cambian dependiendo de la tecnología a utilizar [40]. La precisión de estas tecnologías se puede medir mediante la precisión de la lectura raw (precisión al leer un sólo fragmento de ácido nucleico a la vez) o la precisión de los ensamblajes mediante consensos (reconstrucción de genomas completos)

El umbral para distinguir especies bacterianas utilizando el gen 16S rRNA es de mínimo 97 % de similitud [41] LEEER https://www.microbiologyresearch.org/content/journal/ijsem/10.1099/ijs.0.059774-0

diversidad, riquesa y composición de la comunidad microbiana

Sanger

Incialmente el gen 16S completo se seucenciaba mediante sanger [42] pero debido a su alto costo y complejidad de llevar a cabo en laboratorio

Illumina

como funciona illumina Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms Las tecnologías de secuenciación de segunda generación, como Illumina e IonTorrent permiten secuenciar solo una parte del gen 16S rRNA, debido a que las lecturas son cortas

(300-400pb) [43]. Esto conlleva a que se deba decidir que región del gen secuenciar, para poder obtener la mayor diversidad posible presente en la muestra. Diversos estudios se han realizado para analizar que región permite obtener esta diversidad de manera precisa [44, 45], llegando a descubrirse que ciertos grupos taxonómicos se identifican de mejor manera al utilizar cierta regiones hipervariables [46, 47].

Illumina se ha convertido hoy en día en el estándar más utilizado para la secuenciación del gen 16S debido a su bajo costo, alta precisión (99.9 %) y througput [48]. Sin embargo, esta tecnología esta limitada por el largo de las lecturas, debido a que al secuenciar con illumina solo se puede secuenciar una o dos regiones del gen 16S (de las 9 regiones hipervariables), lo que límita la resolución taxónomica, permitiendo asignar correctamente solo a nivel de genero [49]

Existen diferentes metodologías pensadas para trabajar con illumina,debido a que se sabe que existe presencia de ruido, como la eliminación de quimeras, la eliminación de singleton y otus raros son recomendasos [50, 51], realizar denoising con dada [52] y unoise [53], pero todas estas metodologías no esgtan disponibles para usarse con annopore, debido al porcentaje de error y al secuenciar todo el gen 16s. Mientras algunos estudios han demostrado, que Nanopore presenta un porcentaje de ruido muy bajo, casi nulo [54], y a su vez, al secuenciar el gen completo, y no contar con OTUs, o ASV, permite mejorar el rendimiento de los estimadores de riqueza que se basan en estas metodologias, como los índices de chao1 [55], ACE [56], o brakaway [57]

Illumina sobrerepresenta el numero de especies debido al ruido? Nanopore is preferable over ...

Es por esto, que las tecnologías de tercera generación como Oxford Nanopore y PacBio se presentan como opciones prometedoras para la secuenciación del gen 16S rRNA, debido a su bajo costo y su capacidad de secuenciar el gen completo en una sola lectura, lo que permite una mayor resolución taxonómica a nivel de especie o incluso de strain [54] [58, 59]. revisar estas citas

Oxford Nanopore

Debido a su capacidad de secuenciar lecturas desde unas pocas kilobases hasta megabases [60], permite obtener información genómica más completa y contigua que las tecnologías de secuenciación de segunda generación, como Illumina.

Su bajo costo y su portabilidad, que permiten secuenciar

En sus inicios, la mayor limitante de utilizar Oxford Nanopore era su alto porcentaje de error. En el 2019 estudios reportaban que llegar a una resolución de especie no era aún posible con Nanopore [61]. Estudios posteriores, determinaron que la precisión de secuenciación se encontraba entre el 92 % al cerca del 96 % [58, 59], siendo aún inviable para la asignación taxonómica a nivel de especie creo que deberia tener cuidado con esto, ya que este paper lo dice, pero otros no. Sin embargo, con la introducción de la nueva química a finales del 2021, el porcentaje de error reportado por nanopore disminuye notablemente, llegando a un 99.9 % de precisión, permitiendo la resolución taxónomica de especie [62].buscar más citas

2.2 | Herramientas

Asignación taxónomica

Existen diferentes herramientas para hacer asignación taxonómica de secuencias, algunas incluyen una asignación taxonómica directa a los datos luego del control de calidad, mientras otras herramientas buscan minimizar el error de Oxford Nanopore mediante metodologías de clustering o de algoritmos de maximización de expectativas. Algunas de las más utilizadas para datos de Oxford Nanopore se presentan a continuación:

es posible ponerle numeritos a las subsubsection?

Hoy en día no hay establecidas buenas practicas para el procesamiento del gen 16S rRNA secuenciado mediante Oxford Nanopore, tanto al hablar de la herramienta para hacer asignación taxonómica, como al hablar de la base de datos al utilizar.leer para ver si citar https://academic.oup.com/femsec/article/97/3/fiab001/6098400

Epi2me

Plataforma desarrollada por Oxford Nanopore para el análisis de datos secuenciación obtenidos mediante sus dispositivos. Integra flujos de trabajo para realizar basecalling y demultiplexación, alineamiento, ensamblaje de SARS-CoV-2, asignación taxonómica de gen 16S, 18S, ITS, y metagenómca, variant calling, entre otros.

Mediante la interfaz gráfica el usuario puede seleccionar el análisis a realizar y configurar los parámetros. Debido a su interfaz de fácil uso permite al usuario abstraerse de la ejecución de herramientas o flujos de trabajo y de la necesidad de contar con recursos computacionales para la ejecución de los mismos. Los resultados se pueden descargar y visualizar mediante la misma plataforma.

Para la asignación taxónomica del gen 16S utiliza la herramienta blast con la base de datos de Genbank.

El output de esta herramienta es un archivo en formato CSV con la información de la lectura, asignación taxonómica a nivel de especie, porcentaje de identidad de la asignación, entre otras.

NanoCLUST

Nanoclust [63] es un flujo de trabajo desarrollado en Nextflow para la clasificación de amplicones del gen 16s obtenidos mediante secuenciación de Oxford Nanopore. Incluye pasos previos a la asignación taxonomica, como el basecalling, demultiplexación y control de calidad. Destaca por utilizar un clustering no supervisado (UMAP) y un paso exhaustivo de corrección de lecturas basada en los clusters obtenidos previo a la asignación taxonómica. Utiliza la base de datos de Genbank para realizar la asignación taxonómica.

Cabe destacar que este flujo de trabajo se encuentra descontinuado ya que fue desarrollado utiilzando Nextflow DSL1 (estándar deprecado en la version 22.10.x). Además, debido a que la herramienta ha dejado de recibir soporte por parte de los desarrolladores, no se han actualizado pasos claves, como el basecalling y demultiplexación (pasos opcionales).

El output de esta herramienta es un archivo csv por cada categoría taxonómica (filo, clase, orden, familia, género, especie) con la cantidad de lecturas asignadas a cada taxonomía. De igual forma,

se generan graficos de barra con las asignaciones, y un gráfico de la separación de los clusters. mejorar.

NanoRTax

NanoRTax [64] es un flujo de trabajo desarrollado en Nextflow que cuenta con una interfaz web que permite al usuario visualizar el progreso y resultados del pipeline. Recibe como entrada los archivos FASTQ, a los cuales se les hace un control de calidad mediante fastp, y a continuación se realiza la asignación taxonómica mediante las herramientas Kraken2, Centrifuge y BLAST.

Al igual que NanoCLUST, NanoRTax utiliza DSL1 por lo que no es comptabile con versiones nuevas de Nextflow.

El output de esta herramienta XX

EMU

EMU [65] busca realizar una corrección de errores y mejorar el error de Oxford Nanopore mediante un enfoque basado en algoritmos de maximización de expectativas para generar perfiles taxonómicos de la comunidad microbiona. Permite realizar estos perfiles utilizando diferentes bases de datos, como, la base de datos de Genbank, RDP y Silva v.138. En el caso de realizar analisis de la región ITS, permite integrar las base de datos de UNITE de fungi y eucariotas.

El output de esta herramienta es un archivo en formato TSV con los perfiles taxonómicos encontrados en cada muestra, es decir, el identificador del taxón, abundancia, especie y la información de todas las categorías taxonómicas.

EzBioCloud Microbial Taxonomic Profiling (MTP) pipeline and the PKSSU4.0 database

En algunos estudios se ha utilizado

VSEARCH [35] against the EzBioCloud 16S database.?

Herramientas bioinformáticas

Existen diferentes herramientas bioinformáticas que se pueden utilizar para el análisis y manipulación de datos de secuenciación, a continuación se presentan algunas de las más relevantes para este trabajo:

Guppy

Guppy es una suite de herramientas provista por Oxford Nanopore para realizar procesamientos de datos de secuenciación básicos. Permite realizar basecalling y demultiplexación, alineamiento, detección de bases modificadas, etc.

FastQC

FastQC[66] permite visualizar la calidad de los datos mediante métricas estándar de calidad, contenido GC, distribución de tamaños, niveles de duplicación y contenido de adaptadores.

Genera un reporte en formato html de fácil visualización separado por módulos, donde cada módulo presenta un estado de Aprobado, Fallido o Advertencia (dependiendo de la calidad de los datos). Se desarrollo pensando en tecnología de secuenciación de lecturas cortas, las cuales poseen un porcentaje de error mucho más bajo que las tecnologías de secuenciación de tercera generación y en análisis de genoma completo, por lo que algunos módulos pueden mostrarse como fallidos debido a la naturaleza de los datos de Oxford Nanopore, sin ser datos de baja calidad.

NanoPlot

NanoPlot [67] es una herramienta para la evaluación de calidad de datos de secuenciación de lecturas largas, permite visualizar la información de calidad, largo de lecturas y distribución de estas mediantes gráficos interactivos.

Genera un reporte en formato html y gráficos interáctivos que permiten visualizar la calidad de los datos, longitud de las lecturas, distribución de la calidad y longitud, entre otros.

Fastp

Fastp[68] es una herramienta de alto rendimiento diseñada para el procesamiento de archivos fastq, permite realizar filtrado de secuencias (por calidad, largo), recortar extremos de baja calidad, recortar adaptadores, eliminar colas polyA, etc.

MultiQC

MultiQC [69] es una herramienta que permite resumir la información obtenida por diferentes herramientas bioinformaticas en un solo informe final. También permite integrar varias muestras en un solo reporte, y multiples pasos de analisis en un solo archivo html.

PICRUSt2

PICRUSt2 [70] es una herramienta para la predicción funcional utilizando secuencias marcadoras de genes. Generalmente se utiliza el gen 16S rRNApara realizar la predicción, pero también se pueden usar otros genes marcadores.

El output entrega archivos en formato CSV con la abundancia de los genes ortologos, la clasificación de las enzimas y las vías metabolicas predichos en cada muestra.

LEfSe

LEfSe (Linear discriminant analysis Effect Size) [71] determina las caracteristicas que permiten explicar las diferencias entre diferentes clases o grupos al combinar pruebas estándar de significancia estadistica junto con pruebas que codifican la consistencia biologica y relevancia del efecto encontrado.

vegan package

Vegan [72] es un paquete desarrollado para R que permite realizar análisis de la ecología comunitaria descriptiva. Contiene funciones de análisis de diversidad, metodos de ordenación comunitaria, análisis de disimilitud, funciones para vegetación y ecologos comunitarios.

Taxonkit

Taxonkit [73] permite la manipulación de información taxónomica de NCBI de una manera comprensiva y eficiente. Dado un identificador taxonómico o un nombre de especie se puede obtener el linage completo de esta.

csvtk

csvtk es una herramienta multiplataforma, eficiente y practica para la manipulación de archivos en formato CSV y TSV. Esta herramienta esta desarrollada para utilizarse en conjunto con otras suites de herramientas como TaxonKit, permitiendo obtener resultados de taxonómia de fácil visualización y manipulación para la integración en flujos de trabajo o scripts.

NCBI database

Tanto EPI2ME como Nanoclust utilizan la base de datos de ncbi.

Lenguajes de programación y Frameworks

Nextflow

Nextflow [74] es un framework open source para el desarrollo de flujos de trabajo, el cual permite la ejecución de éstos en diferentes entornos computacionales, ya sea en un computador personal, una plataforma de cómputo de alto rendimiento o en la nube. También permite la ejecución de flujos de trabajo de manera paralela, manejando los recursos computacionales de manera eficiente, y sencilla para el usuario. Al permitir el desarrollo de flujos de trabajo escalables y reproducibles es una buena alternativa que ha ganado popularidad debido a su facilidad de uso.

Cuenta con una comunidad llamada nf-core que se encarga de desarrollar flujos de trabajo para el análisis de datos biologicos, los cuales son revisados por la comunidad y publicados en su repositorio. Esto permite contar con una gran cantidad de flujos de trabajo disponibles, los cuales pueden ser ejecutados de manera sencilla por los usuarios, pero cabe destacar que hay que tener conocimientos de linea de comando para poder ejecutarlos.

python?

js

r?

FastAPI

Framework rápido y ligero para el desarrollo de APIs modernas de manera ágil utilizando Python y basado en sus anotaciones de tipo estandar. Utiliza pydantic para la validación de los datos de entrada y salida y starlette para el manejo de las peticiones HTTP. no estoy 100 % segura

SQLalchemy

ORM (Object-Relational Mapping) para Python que permite la comunicación entre el back end y la base de datos SQL de manera sencilla, transformando los resultados de la base de datos en

estructuras utilizables mediante Python. Gestiona la creación de modelos y consultas de forma sencilla.

Vue.js

Vue es un framework para la construcción de interfaces de usuario. Se basa en JavaScript, HTML y CSS para proporcionar un modelo de programación declarativo y basado en componentes que permite desarrollar interfaces de manera eficiente.

TypeScript

TypeScript [40] es un lenguaje de programación basado en JavaScript, el cual añade sintaxis adicional a JavaScript (o frameworks basados en JS) para soportar la integración de tipado de datos. Al especificar los tipos de datos, TypeScript tiene la capacidad de validarlos e informar errores cuando estos no correspondan.

Vuetify

Vuetify es un proyecto de código abierto para la construcción de interfaces utilizando los componentes de Vue. Permite la personalización de los componentes con SASS y SCSS, cuenta con un diseño responsivo, y una gran cantidad de componentes ya predefinidos.

PostgreSQL

PostgreSQL es un sistema de gestión de bases de datos relacionales de código abierto basado en POSTGRES. Permite el uso de tipos de datos complejos realizar consultas tanto relacionales (SQL) y no relacionales (JSON).

Gestores de paquetes

Conda

Conda [75] es una herramienta de código abierto, multiplataforma que permite la gestión de paquetes, dependencias y entornos de desarrollo de manera sencilla. Permite aislar entornos virtuales con caracteristicas especificas, lo que facilita la reproducibilidad de los análisis y la portabilidad de los mismos. já

Apptainer

Apptainer (antes llamado Singularity [76]) simplifica la creación y ejecución de contenedores, asegurando el encapsulamiento de los componentes de softwares necesarios para su reproducibilidad y portabilidad.

Docker

Creo que no es necesario

Métricas para la evaluación de la diversidad microbiana

La diversidad microbiana

Los índices de diversidad se dividen en índices de riqueza o uniformidad/divergencia

Indice de simpson

Indice de shannon

Indice de chao2

3

Flujo de trabajo y Aplicación Web

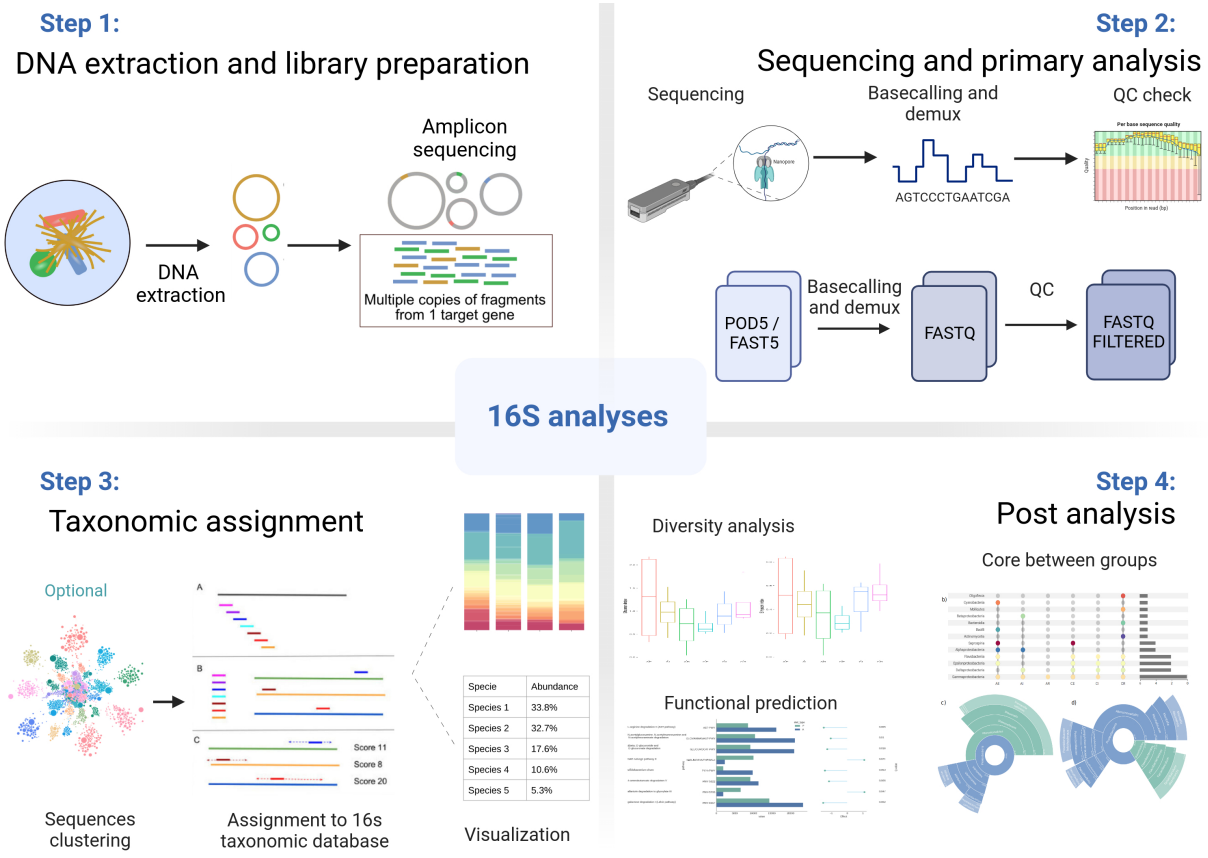


Figura 3.1. Flujo de trabajo estándar para la secuenciación y análisis de secuenciaciones del gen 16S

3.1 | Flujo de trabajo

Se desarrolló un flujo de trabajo automatizado en Nextflow que permite el análisis y caracterización de secuencias del gen 16S obtenidas mediante Oxford Nanopore. El flujo de trabajo está desarrollo de manera modular permitiendo que el usuario pueda personalizar su ejecución, agregando o quitando subworkflows o módulos según sus necesidades. Cuenta con los siguientes módulos:

- Basecalling y demultiplexacion:
- Control de calidad:
- Asignación taxonómicas:
- Caracterización de la comunidad microbiana:

Desarrollo de contenedores o conda enviroments

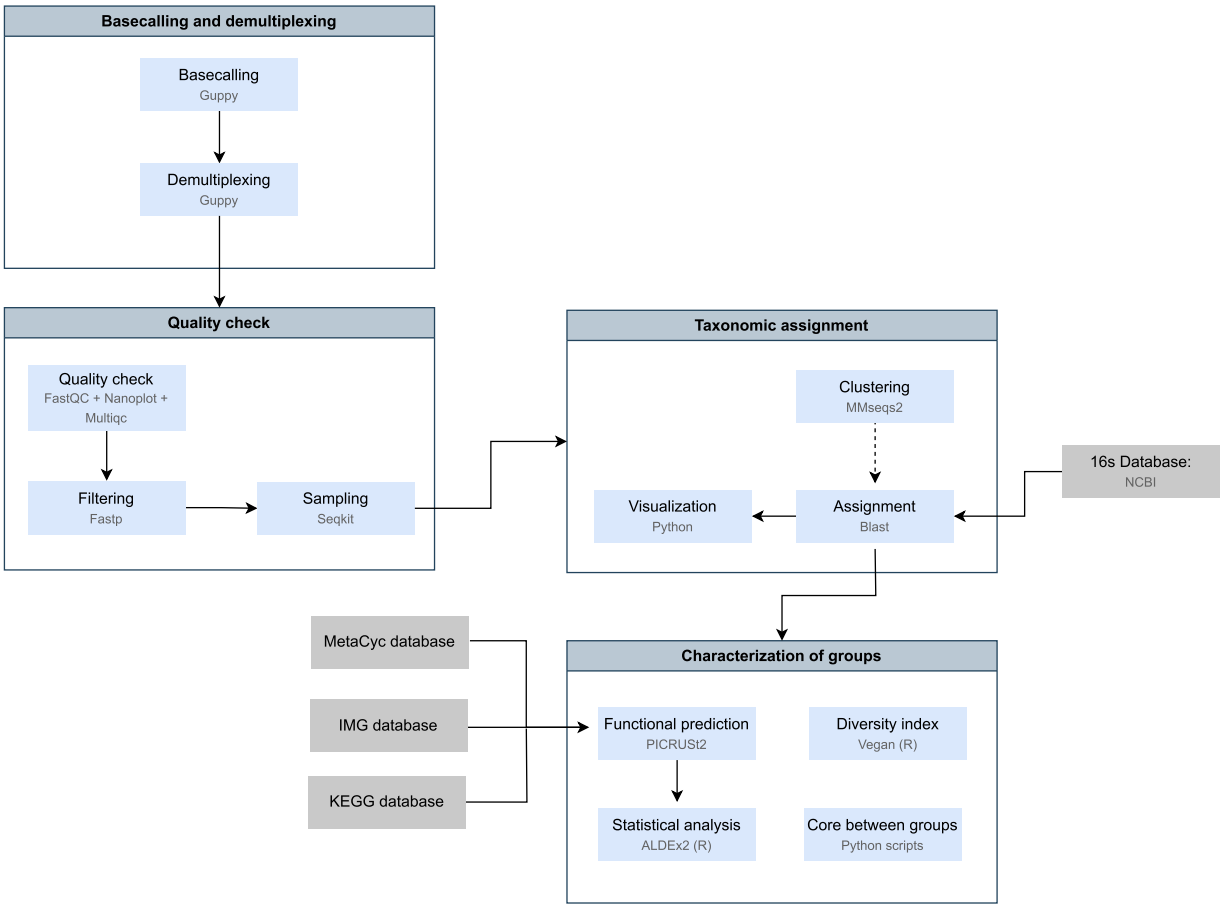


Figura 3.2. Bioinformatics pipeline

El flujo de trabajo cuenta con parámetros generales que son independientes de los módulos:

- stage: Subworkflow a ejecutar. Puede contener los siguientes valores:
  - All_without_basecalling: Ejecutar todo el flujo de trabajo sin incluir basecalling y demultiplexación.

    - All_with_basecalling: Ejecutar todo el flujo de trabajo incluyendo el proceso de basecalling y demultiplexacion.
    - Preprocessing: Ejecutar solo el subworflow de preprocesamiento
    - TaxonomyAssignament: Ejecutar el modulo de asignación taxonomca.
    - Visualization: Ejecutar solo el subworkflow de visualización de gráficos
    - Functional: Ejecutar solo el subworkflow de predicion funcional
    - Diversity: Ejecutar solo los modulos de indices de diversidad
- input:
- samples.csv:

Ejecución de flujo de trabajo completo

hacer un esquema de los casos del uso, que el usuario puede comenzar con fastq, pod5, o hacer solo los graficos, etc

Control de calidad

Este módulo primero se encarga de realizar un control de calidad a las secuencias mediante las herramientas FastQC y Nanoplot. Posteriormente, se eliminan secuencias de baja calidad y secuencias no pertenecientes al gen 16s (basado en el tamaño de la lectura). Para esto, se utiliza el software Fastp [] que permite realizar las tareas mencionadas anteriormente.

Input: Output: Parámetros:

FastQC y Nanoplot

FastQC y Nanoplot reciben como input el archivo FASTQ a procesar y entrega como output un archivo HTML con el reporte del control de calidad. Este reporte será utilizado para integrar la información de este proceso en el reporte de MultiQC. En el flujo de trabajo ambas herramientas se utilizarán dos veces, una vez con el archivo FASTQ raw y otra vez con el archivo FASTQ preprocesado.

Fastp Parametros

- q 15
- cut_mean_quality 15
- length_required 1000
- length_limit 2200
- disable_adapter_trimming
- disable_trim_poly_g

Fastp recibe como input el archivo FASTQ raw y entrega como output el archivo FASTQ preprocesado. De igual manera, entrega un reporte en formato HTML y un reporte en formato JSON, el cual posteriormente será utilizado para integrar la información de este proceso en el reporte de MultiQC.

SEQKIT

Seqkit se utilizará para realizar un subsampleo de la muestra y evitar sesgos a la hora de realizar la caracterización por grupos debido a la cantidad de lecturas de cada muestra. Recibe como input el archivo FASTQ preprocesado. Posteriormente el archivo subsampleado se transformará a formato FASTA mediante la misma herramienta, y será utilizado para realizar la caracterización de la comunidad microbiana.

Asignación taxónomica

Clustering con MMSEQS

BLAST

Para la asignación taxonómica se utiliza BLAST con la base de datos de 16s de Genbank.

- -mt_mode 1
- -outfmt ”6 qseqid staxids sscinames evalue length pident qcovs”
- -evalue 0.001
- -max_target_seqs 1

Para asegurar que las secuencias sean asignadas a la taxonomía correcta, se utiliza un evalue de 0.001 y un porcentaje de identidad mayor a XX EL output de BLAST es un archivo de formato tabular (TSV) que contiene el identificador de la muestra, el id de la taxonomia asociada a la secuencia, el nombre científico de la taxonomía, el valor de evalue, el largo de la secuencia, el porcentaje de identidad y el porcentaje de cobertura de la secuencia.

Mediante TaxonKit se obtiene el linage completo de la taxonomía asignada a la secuencia , y se formatea mediante CSVTk.

MERGE_BLAST_OUT

Se desarrollo un script en python para resumir la asignación taxonomica obtenida con BLAST de todas las muestras en un solo archivo. Además, se generan archivos por cantidad de lecturas y porcentaje, y por cada categoria taxonomica.

Caracterización de la comunidad microbiana

Stacked Plots

Se desarrollo un script en python que grafica las taxonomias en graficos de barras apiladas, permitiendo visualizar la distribución de las taxonomias en las muestras.En caso de que el usuario haya ingresado grupos asignados a cada muestra también se obtendran graficos de barras apiladas de los grupos. Las taxonomías menores a un 1 % (valor por defecto y modificable) se agrupan en una nueva taxonomia llama .otrosz se muestra en color gris.

- –taxonomy_input_file: Archivo en formato csv donde las columnas son las muestras, las filas las taxonomicas y el valor de abundancia de cada taxonomia en cada muestra en el centro.

- –annotate_tax: En caso de ser true se anotara el nombre de la taxonomia y el porcentaje en la barra.
- min_abundance_annotate: Porcentaje de abundancia maximo para agrupar las taxonomicas en sección de otros

Core Plot

Se desarrollo un script en python que realiza un grafico circular jerarquico de las taxonomias compartidas entre las muestras. Este grafico puede realizarse a nivel de grupos (siempre que el usuario hubiera ingresado los grupos en el archivo de metadata), y también se realiza entre todas las muestras. Input: CSV de taxonomia a nivel de especie por muestra y archivo de linajes. Output: Grafico circular jerarquico de las taxonomias compartidas entre las muestras en formato PDF

Indices de diversidad

shannon simpson chao

Prediccion Funcional

La predicción funcional se realiza mediante la herramienta PICRUSt2.

Input: output: Picrust2 + aldex + Lefse

3.2 | Base de datos

Tabla de Usuarios

campo	tipo de dato	descripción	PK	FK
id	int	ID	Si	No
username	str	Nombre de usuario	No	No
fullname	str	Nombre completo	No	No
email	str	Email	No	No
hasshed_password	str		No	No
is_active	bool		No	No
role_platform_id	int		No	Si

Tabla 3.1. Tabla Users

Roles de la plataforma

campo	tipo de dato	descripción	PK	FK
id	int	ID	Si	No
name	str	Nombre del rol (admin / basic user)	No	No
description	str	Descripción del rol	No	No

Tabla 3.2. Tabla PlatformRole

Projectos

campo	tipo de dato	descripción	FK
id	int (PK)	ID	No
name	str	Nombre del projecto	No
description	str	Descripción del proyecto	No
samples	JSON/dict[str,any}	Metadata de las muestras. Puede contener como claves: Barcode, Sample, Group, Subgroup	No
colaboration	str	???	No
type_file_in	str	Tipo de archivo a procesar (POD5, FASTQ, CSV, FASTA)	No
path_data	str	Ruta en el servidor donde se almacenaran los archivos de entrada	No
status	str	Estado del proyecto en ejecución (upload_data / running / failed / finish)	No
note	JSON/dict[str,any}	Información sobre el proyecto (cantidad de muestras total, procesadas y descartadas)	No
logs	JSON/dict[str,any}	Logs	No
owner_id	int	Id del usuario que sube el proyecto	Si

Tabla 3.3. Tabla de Proyectos

Roles del Proyecto

campo	tipo de dato	descripción	FK
id	int (PK)	ID	No
name	str	Nombre del rol (leer, escribir, eliminar)	No
description	str	Descripción del rol	No

Tabla 3.4. Tabla de Roles de Proyectos

Roles Projectos y Usuarios

campo	tipo de dato	descripción	FK
user_id	int (PK)	ID del usuario	Si
project_id	int (PK)	ID del proyecto	Si
role_id	int (PK)	Rol del usuario dentro del proyecto en especifico	Si

Tabla 3.5. Tabla de Roles Proyectos y Usuarios

Resultados

campo	tipo de dato	descripción	FK
id	int	ID	No
taxonomic_assignament	JSON/dict[str,any]	Assignación taxonomica por muestra y por nivel taxonómico (especie, genero, familia, orden, clase y filo)	No
functional_prediction	JSON/dict[str,any]	Resultados de la predicción funcional por muestra y por categoría funcional (EC, Pathways, KO)	No
core	JSON/dict[str,any]	Cantidad de lecturas/porcentaje compartido entre los grupos	No
basic_statistics	JSON/dict[str,any]	Estadísticas básicas de las muestras (Total de lecturas procesadas y sin procesar, calidad y largo promedio)	No
diversity	JSON/dict[str,any]	Calculo de los indices de diversidad (Shannon, Simpson, Chao2)	No
config_run	JSON/dict[str,any]	Parámetros personalizados por el usuario	No
project_id	int	ID del proyecto	Si

Tabla 3.6. Tabla de Resultados

3.3 | Aplicación Web

Se desarrollo una aplicación web mediante Vue3 y FastAPI que permite al usuario subir sus archivos de secuenciación y metadata. Con esto el usuario puede abstraerse de tener conocimiento en línea de comando o ejecución de herrsamientas bioinformaticas o flujos de trabajo, ya que mediante la interfaz web el usuario selecciona los análisis que desea realizar. La información ingresada por el usuario es guardada en la base de datos y la misma plataforma/bd/script se encarga de ejecutar el pipeline de manera automatica. Una vez el pipeline termina de ejecutarse se escriben todos los resultados (o durante la ejecución) en la base de datos. La plataforma lee esta información en la base de datos y despliega los resultados en el menu de proyectos y analisis.

Middleware

Security

Login

Al ingresar en la página de Login, el usuario deberá ingresar su nombre de usuario y contraseña. En caso de que los datos sean correctos será redireccionado a la página de proyectos (Ver sección siguiente.). En caso de que los datos sean incorrectos se mostrará un mensaje de error con el mensaje *"Usuario o contraseña incorrectos"* y deberá ingresar sus credenciales nuevamente. Debo añadir que la gente pueda crear su propia cuenta D;

Cambio de contraseña

En el caso de que el usuario desee cambiar su contraseña puede hacerlo en esta sección. Para ello deberá escribir su contraseña actual y su nueva contraseña dos veces. En caso de que la contraseña actual sea incorrecta se mostrará un mensaje de error con el mensaje *"Contraseña actual incorrecta"*. En caso de que las contraseñas nuevas no coincidan se mostrará un mensaje de error con el mensaje *"Las contraseñas no coinciden"*. En caso de que la contraseña sea cambiada con éxito se mostrará un mensaje de éxito con el mensaje *"Contraseña cambiada con exito"*.

Resultados/Proyectos

Una vez que el usuario valida sus credenciales en la plataforma será redireccionado a la sección de Resultados. En esta sección se mostraran los proyectos que el usuario ha subido a la plataforma, estos proyectos pueden estar en ejecución, finalizados o finalizados con errores. Por cada proyecto se desplegará la información básica en una tarjeta:

- Nombre del proyecto
- Descripción del proyecto
- Cantidad de muestras procesadas, descartadas y totales
- Estado del proyecto (upload_data / running / failed / finish)
- En caso de que el proyecto haya finalizado el usuario podrá acceder a la sección especifica de resultados del proyecto mediante el botón de Ver resultados.

Nuevo análisis

En esta sección el usuario podrá ingresar información de un proyecto de secuenciación al cual se le desea realizar los análisis.

El usuario deberá rellenar la información básica del proyecto como, Nombre del proyecto y descripción. El usuario puede subir la información en diferentes formatos, dependiendo del análisis que desee realizar.

- POD5: En caso de querer comenzar desde el basecalling.
- FASTQ: En caso de querer saltarse el paso de basecalling y demultiplexación e iniciar directamente con el control de calidad o asignación taxonómica.

A continuación el usuario deberá subir un archivo de metadata en formato (XLXS) que deberá contener la información de la muestras:

- sample: identificador de la muestra (obligatorio)
- barcode: barcode que identifica la muestra (en caso de querer realizar basecalling demultiplexación)
- group: groupo al que pertenece cada muestra (en caso de querer hacer diferenciación entre grupos)
- subgroup: subgrupo al que pertenece cada muestra (en caso de querer hacer diferenciación entre subgrupos)

A continuación el usuario deberá seleccionar los análisis que desea realizar. Cabe destacar que el usuario puede seleccionar todos los análisis o solo algunos de ellos. Los análisis disponibles son:

- Basecalling y demultiplexacion
- Filtros y control de calidad
- Asignación taxonomica

- Indices de diversidad
- Predicción funcional

La plataforma se encarga de verificar si el archivo de metadata cuenta con la información necesaria para realizar cada analisis. En caso de que el archivo de metadata no cuente con la información necesaria y el usuario desee realizar uno de esos analisis, se desplegara un mensaje de error al lado del análisis indicando que información se debe añadir en el archivo de metadata. Los parámetros que se pueden modificar son los siguientes:

- Basecalling y demultiplexacion: Flowcell, kit de ligación y kit de barcoding utilizados durante la secuenciación. Modelo a utilizar para realizar el basecalling
- QC: Longitud mínima y máxima en pares de bases de las lecturas, calidad mínima de las lecturas y cantidad de lecturas a utilizar para los análisis posteriores(subsampleo).
- Predicción funcional: completar

En la parte derecha del componente el usuario puede visualizar los parámetros por defecto y modificarlos en caso de que lo desee.

En la parte inferior del componente se encuentra el botón de Subir data, el cual al hacer click en el, ingresará la información a la base de datos y copiará los archivos al cluster/servidor. Una vez que el usuario presionar el botón de subir data, la plataforma se encarga de verificar que se cuente con toda la información necesaria para correr el pipeline.

Resultados de un proyecto en especifico

Una vez que el pipeline haya finalizado su ejecución, la plataforma leerá los resultados desde la base de datos y desplegará la información en la sección de resultados de un proyecto en especifico. Esta sección cuenta con 5 subsecciones, cada una con información específica del análisis realizado. En caso de que al ingresar el proyecto el usuario no seleccione todos los análisis a ejecutar, solo se mostrarán las secciones indicadas por el usuario.

De que muestras puedo poner información en mi tesis?

Informacion Basica de las muestras

Esta sección se mostrará siempre que el usuario empiece con archivos POD5 o FastQ, es decir, ya sea comenzando el análisis desde el basecalling o desde el control de calidad. Igual si es que solo se hace asignaicón taxonomica.

Al lado izquiero del componente se puede visualizar una tabla con la información básica de cada muestra:

- Nombre de la muestra
- Total de lecturas
- Calidad promedio
- Largo promedio

- Total de lecturas después del control de calidad
- Nota (en caso de que la muestra haya sido descartada se indica en esta columna)

En la parte inferior de la tabla hay una nota que indica la cantidad de lecturas que se consideraron para los análisis posteriores (subsampleo).

En la parte derecha de la sección se puede visualizar un heatmap donde en el eje X se encuentra el tamaño de las lecturas, y en el eje Y la calidad. El color indica la cantidad de lecturas que se encuentran en esa intersección, mientras más azul, más lecturas tienen la calidad y tamaño indicado. Para este grafico se consideraron todas las muestras con sus lecturas antes de los filtros de calidad o después?.

En la parte inferior del gráfico hay una nota que indica la cantidad de lecturas que se encuentran en un rango especifico. Este valor es variable .....

Asignación taxonómica

La sección de assignación taxonómica cuenta con un gráfico de barras apiladas y una tabla por cada categoría taxonómica (especie, genero, familia, orden, clase y filo). En la leyenda del gráfico de barras apiladas solo se muestran las 10 taxonomías con mayor abundancia. El usuario puede moverse através del gráfico y posicionarse en las barras para poder visualizar el nombre de la taxonomía y el porcentaje o cantidad de lecturas en una muestra en especifico. Todas aquellas taxonomías que tienen un porcentaje menor a un 1 % (valor por defecto) son agrupadas en una nueva taxonomia llamada *Otros*. En la parte superior de la tabla se encuentra un campo de texto donde el usuario puede buscar una taxonomía en especifico y visualizar su abundancia o cantidad de lecturas en todas las muestras.

El usuario puede visualizar la información en porcentaje o en cantidad de lecturas manipulando el boton de la parte inferior del gráfico. Asi también puede modificar el porcentaje utilizado para generar la taxonomía Otros.

En caso de que el usuario al crear el proyecto hubiera ingresado información de los grupos asociados a las muestras, se podrá visualizar un nuevo grupo de botones que permiten al usuario visualizar la información de las taxonomías por grupo o por muestra.

Similitud entre las muestras

En esta sección se puede visualizar un gráfico Sunburst el cual representa las taxonomías compartidas entre una cantidad de muestras especificas. Mientras más grande el anillo en el gráfico, indica que la presencia de esa taxonomía es mayor. Cada nivel de anillo representa una categoria taxonomica, siendo la más interna especie y la más externa filo. En caso de que el usuario haya ingresado grupos asociados a las muestras, se podrán visualizar más gráficos de Sunburst en la sección, un gráfico por cada grupo. En el caso de que el usuario no haya ingresado grupos, podrá visualizar solamente un gráfico que mostrará las taxomías compartidas entre todas las muestras.

En caso de que un grupo no tenga taxonomías compartidas entre si, se desplegará un mensaje indicando esto. En la parte inferior del gráfico, al lado derecho de la leyenda se puede visualizar un icono, el cual al posicionarse sobre el, va a mostrar las muestras utilizadas para generar el gráfico.

Indices de diversidad

En caso de que el usuario hubiera ingresado grupos al inicio del proyecto, se podrá visualizar tres gráficos boxplot, uno por cada indice de diversidad (Shannon, Simpson y Chao2). En caso de que el usuario no haya ingresado grupos, esta sección no se desplegará en la plataforma.

Predicción funcional

En esta sección se presenta una tabla con la información de la predicción funcional obtenida mediante PICRUSt2 (EC, KO y Pathways) para cada muestra. En la parte superior se puede visualizar un campo de texto de búsqueda con el cual el usuario puede filtrar la información de la tabla.

En el lado derecho de la sección, en caso de que el usuario hubiera ingresado grupos, se puede ver un gráfico de barras horizontales que muestra los pathways con diferencias significativas entre los grupos (información obtenida mediante Lefse). En caso de que no se haya ingresado información de grupos, solo se desplegará la tabla.

Descarga de los resultados

En la parte superior derecha de la sección de resultados, se encuentra un botón con el texto Download data. Al hacer click en este botón se descargará un archivo comprimido con toda los resultados generados por el pipeline.

- CSV de asignación taxonomica por muestra y por grupo (en caso de ingresarse), y por porcentaje y cantidad de lecturas
- CSV de predicción funcional (EC, KO y pathways)
- CSV con los valores del calculo de los indices de diversidad
- PDF con los gráficos de barras apiladas, Sunburst y boxplot
- Archivo de texto con la información del pipeline (versión, parámetros, etc)

graficos, csv, información de las bds y versión del pipeline

Documentación

En esta sección se despliega la documentación del pipeline, la cual cuenta con la información de los módulos, parámetros y herramientas utilizadas en el pipeline. La documentación se encuentra dividida por cada módulo, donde se muestra la versión de la herramienta utilizada y los parámetros por defecto y modificables por el usuario.

Basecalling y demultiplexación

Control de calidad

Asignación taxonómica

Predicción funcional

Indices de diversidad

4

Discusión

glosario

- lecturas
- phread
- pares de bases
- API (application programming interface o interfaz de programaciónn de aplicaciones),
- JSON
- ORM (Object-Relational Mapping)
- UMAP
- FASTQ
- POD5
- FASTA
- CSV
- evalue: The statistical significance threshold for reporting matches against database sequences
- min coverage: Minimum horizontal coverage for a query sequence to be considered a match
- Min identity: Minimum proportion of identical bases between the query and the subject sequence
- Max target sequences: Number of aligned sequences to keep for each query
- througput
- raw
- profundidad: Multiples lecturas en una misma región

Bibliografía

1. Lamport, L. *LaTeX: A Document Preparation System* (Addison-Wesley, Reading, Massachusetts, 1994).
2. Gilbert, J. A., Blaser, M. J., *et al.* Current understanding of the human microbiome. *Nature medicine* **24,** 392–400 (2018).
3. Bahrndorff, S., Alemu, T., Alemneh, T., Lund Nielsen, J., *et al.* The microbiome of animals: implications for conservation biology. *International journal of genomics* **2016** (2016).
4. Berendsen, R. L., Pieterse, C. M. & Bakker, P. A. The rhizosphere microbiome and plant health. *Trends in plant science* **17,** 478–486 (2012).
5. Moran, M. A. The global ocean microbiome. *Science* **350,** aac8455. doi:10.1126/science.aac8455. eprint: https://www.science.org/doi/pdf/10.1126/science.aac8455 (2015).
6. Banerjee, S. & Van Der Heijden, M. G. Soil microbiomes and one health. *Nature Reviews Microbiology* **21,** 6–20 (2023).
7. Marco, M. L. Defining how microorganisms benefit human health. *Microbial Biotechnology* **14,** 35–40 (2021).
8. Fijan, S. Microorganisms with claimed probiotic properties: an overview of recent literature. *International journal of environmental research and public health* **11,** 4745–4767 (2014).
9. Altveş, S., Yildiz, H. K. & Vural, H. C. Interaction of the microbiota with the human body in health and diseases. *Bioscience of microbiota, food and health* **39,** 23–32 (2020).
10. Hou, K., Wu, Z.-X., *et al.* Microbiota in health and diseases. *Signal transduction and targeted therapy* **7,** 1–28 (2022).
11. Ursell, L. K., Clemente, J. C., *et al.* The interpersonal and intrapersonal diversity of human-associated microbiota in key body sites. *Journal of Allergy and Clinical Immunology* **129,** 1204–1208 (2012).
12. Sender, R., Fuchs, S. & Milo, R. Revised estimates for the number of human and bacteria cells in the body. *PLoS biology* **14,** e1002533 (2016).
13. Ley, R. E., Peterson, D. A. & Gordon, J. I. Ecological and evolutionary forces shaping microbial diversity in the human intestine. *Cell* **124,** 837–848 (2006).
14. Bitton, G. Role of microorganisms in biogeochemical cycles. *Wastewater Microbiology.,* 51–73 (1994).
15. Gougoulias, C., Clark, J. M. & Shaw, L. J. The role of soil microbes in the global carbon cycle: tracking the below-ground microbial processing of plant-derived carbon for manipulating carbon dynamics in agricultural systems. *Journal of the Science of Food and Agriculture* **94,** 2362–2371 (2014).
16. Jiao, S., Chen, W. & Wei, G. Linking phylogenetic niche conservatism to soil archaeal biogeography, community assembly and species coexistence. *Global Ecology and Biogeography* **30,** 1488–1501 (2021).
17. Bickel, S. & Or, D. Soil bacterial diversity mediated by microscale aqueous-phase processes across biomes. *Nature Communications* **11,** 116 (2020).
18. Hu, J., Wei, Z., *et al.* Probiotic Pseudomonas communities enhance plant growth and nutrient assimilation via diversity-mediated ecosystem functioning. *Soil Biology and Biochemistry* **113,** 122–129 (2017).
19. Lemanceau, P., Blouin, M., Muller, D. & Moënne-Loccoz, Y. Let the core microbiota be functional. *Trends in Plant Science* **22,** 583–595 (2017).
20. Hardoim, P. R., Van Overbeek, L. S., *et al.* The hidden world within plants: ecological and evolutionary considerations for defining functioning of microbial endophytes. *Microbiology and molecular biology reviews* **79,** 293–320 (2015).
21. Vorholt, J. A. Microbial life in the phyllosphere. *Nature reviews microbiology* **10,** 828–840 (2012).
22. Compant, S., Samad, A., Faist, H. & Sessitsch, A. A review on the plant microbiome: Ecology, functions, and emerging trends in microbial application. *Journal of Advanced Research* **19.** Special Issue on Plant Microbiome, 29–37. doi:https://doi.org/10.1016/j.jare.2019.03.004 (2019).
23. Roslund, M. I., Puhakka, R., *et al.* Biodiversity intervention enhances immune regulation and health-associated commensal microbiota among daycare children. *Science advances* **6,** eaba2578 (2020).
24. Tun, H. M., Konya, T., *et al.* Exposure to household furry pets influences the gut microbiota of infants at 3–4

months following various birth scenarios. *Microbiome* **5,** 1–14 (2017).
25. Azad, M. B., Konya, T., *et al.* Infant gut microbiota and the hygiene hypothesis of allergic disease: impact of household pets and siblings on microbiota composition and diversity. *Allergy, Asthma & Clinical Immunology* **9,** 1–9 (2013).
26. Kates, A. E., Jarrett, O., *et al.* Household pet ownership and the microbial diversity of the human gut microbiota. *Frontiers in cellular and infection microbiology* **10,** 73 (2020).
27. Yan, T., O'brien, P., Shelton, J., Whelen, A. & Pagaling, E. Municipal wastewater as a microbial surveillance platform for enteric diseases: a case study for Salmonella and salmonellosis. *Environmental science & technology* **52,** 4869–4877 (2018).
28. Rackaityte, E. & Lynch, S. V. The human microbiome in the 21st century. *Nature communications* **11,** 5256 (2020).
29. Janda, J. M. & Abbott, S. L. 16S rRNA gene sequencing for bacterial identification in the diagnostic laboratory: pluses, perils, and pitfalls. *Journal of clinical microbiology* **45,** 2761–2764 (2007).
30. López-Aladid, R., Fernández-Barat, L., *et al.* Determining the most accurate 16S rRNA hypervariable region for taxonomic identification from respiratory samples. *Scientific reports* **13,** 3974 (2023).
31. Patel, J. B. 16S rRNA gene sequencing for bacterial pathogen identification in the clinical laboratory. *Molecular diagnosis* **6,** 313–321 (2001).
32. Clarridge III, J. E. Impact of 16S rRNA gene sequence analysis for identification of bacteria on clinical microbiology and infectious diseases. *Clinical microbiology reviews* **17,** 840–862 (2004).
33. Klindworth, A., Pruesse, E., *et al.* Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies. *Nucleic acids research* **41,** e1–e1 (2013).
34. Mizrahi-Man, O., Davenport, E. R. & Gilad, Y. Taxonomic classification of bacterial 16S rRNA genes using short sequencing reads: evaluation of effective study designs. *PloS one* **8,** e53608 (2013).
35. Guo, F., Ju, F., Cai, L. & Zhang, T. Taxonomic precision of different hypervariable regions of 16S rRNA gene and annotation methods for functional bacterial groups in biological wastewater treatment. *PloS one* **8,** e76185 (2013).
36. Soergel, D. A., Dey, N., Knight, R. & Brenner, S. E. Selection of primers for optimal taxonomic classification of environmental 16S rRNA gene sequences. *The ISME journal* **6,** 1440–1444 (2012).
37. Woo, P. C., Lau, S. K., Teng, J. L., Tse, H. & Yuen, K.-Y. Then and now: use of 16S rDNA gene sequencing for bacterial identification and discovery of novel bacteria in clinical microbiology laboratories. *Clinical Microbiology and Infection* **14,** 908–934 (2008).
38. Tanner, A., Maiden, M. F., Paster, B. J. & Dewhirst, F. E. The impact of 16S ribosomal RNA-based phylogeny on the taxonomy of oral bacteria. *Periodontology 2000* **5,** 26–51 (1994).
39. Pollock, J., Glendinning, L., Wisedchanwet, T. & Watson, M. The madness of microbiome: attempting to find consensus "best practice" for 16S microbiome studies. *Applied and environmental microbiology* **84,** e02627–17 (2018).
40. Bierman, G., Abadi, M. & Torgersen, M. *Understanding typescript* in *European Conference on Object-Oriented Programming* (2014), 257–281.
41. Kim, M., Oh, H.-S., Park, S.-C. & Chun, J. Towards a taxonomic coherence between average nucleotide identity and 16S rRNA gene sequence similarity for species demarcation of prokaryotes. *International journal of systematic and evolutionary microbiology* **64,** 346–351 (2014).
42. Sanger, F. & Coulson, A. R. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *Journal of molecular biology* **94,** 441–448 (1975).
43. Salipante, S. J., Kawashima, T., *et al.* Performance comparison of Illumina and ion torrent next-generation sequencing platforms for 16S rRNA-based bacterial community profiling. *Applied and environmental microbiology* **80,** 7583–7591 (2014).
44. Liu, Z., DeSantis, T. Z., Andersen, G. L. & Knight, R. Accurate taxonomy assignments from 16S rRNA sequences produced by highly parallel pyrosequencers. *Nucleic acids research* **36,** e120–e120 (2008).
45. Schloss, P. D., Gevers, D. & Westcott, S. L. Reducing the effects of PCR amplification and sequencing artifacts on 16S rRNA-based studies. *PloS one* **6,** e27310 (2011).
46. He, Y., Zhou, B.-J., *et al.* Comparison of microbial diversity determined with the same variable tag sequence extracted from two different PCR amplicons. *BMC microbiology* **13,** 1–8 (2013).
47. Claesson, M. J., Wang, Q., *et al.* Comparison of two next-generation sequencing technologies for resolving highly complex microbiota composition using tandem variable 16S rRNA gene regions. *Nucleic acids research* **38,** e200–e200 (2010).
48. Pichler, M., Coskun, Ö. K., *et al.* A 16S rRNA gene sequencing and analysis protocol for the Illumina MiniSeq platform. *Microbiologyopen* **7,** e00611 (2018).
49. Johnson, J. S., Spakowicz, D. J., *et al.* Evaluation of 16S rRNA gene sequencing for species and strain-level microbiome analysis. *Nature communications* **10,** 5029 (2019).
50. Caporaso, J. G., Lauber, C. L., *et al.* Global patterns of 16S rRNA diversity at a depth of millions of sequences per

sample. *Proceedings of the national academy of sciences* **108,** 4516–4522 (2011).

51. Auer, L., Mariadassou, M., O'Donohue, M., Klopp, C. & Hernandez-Raquet, G. Analysis of large 16S rRNA Illumina data sets: Impact of singleton read filtering on microbial community description. *Molecular ecology resources* **17,** e122–e132 (2017).
52. Callahan, B. J., McMurdie, P. J., *et al.* DADA2: High-resolution sample inference from Illumina amplicon data. *Nature methods* **13,** 581–583 (2016).
53. Edgar, R. C. UNOISE2: improved error-correction for Illumina 16S and ITS amplicon sequencing. *BioRxiv,* 081257 (2016).
54. Szoboszlay, M., Schramm, L., *et al.* Nanopore is preferable over Illumina for 16S amplicon sequencing of the gut microbiota when species-level taxonomic classification, accurate estimation of richness, or focus on rare taxa is required. *Microorganisms* **11,** 804 (2023).
55. Chao, A. Nonparametric estimation of the number of classes in a population. *Scandinavian Journal of statistics,* 265–270 (1984).
56. Chao, A. & Lee, S.-M. Estimating the number of classes via sample coverage. *Journal of the American statistical Association* **87,** 210–217 (1992).
57. Willis, A. & Bunge, J. Estimating diversity via frequency ratios. *Biometrics* **71,** 1042–1049 (2015).
58. Urban, L., Holzer, A., *et al.* Freshwater monitoring by nanopore sequencing. *Elife* **10,** e61504 (2021).
59. Delahaye, C. & Nicolas, J. Sequencing DNA with nanopores: Troubles and biases. *PloS one* **16,** e0257521 (2021).
60. Amarasinghe, S. L., Su, S., *et al.* Opportunities and challenges in long-read sequencing data analysis. *Genome biology* **21,** 30 (2020).
61. Winand, R., Bogaerts, B., *et al.* Targeting the 16s rRNA gene for bacterial identification in complex mixed samples: Comparative evaluation of second (illumina) and third (oxford nanopore technologies) generation sequencing technologies. *International journal of molecular sciences* **21,** 298 (2019).
62. Yoon, S.-H., Ha, S.-M., *et al.* Introducing EzBioCloud: a taxonomically united database of 16S rRNA gene sequences and whole-genome assemblies. *International journal of systematic and evolutionary microbiology* **67,** 1613–1617 (2017).
63. Rodríguez-Pérez, H., Ciuffreda, L. & Flores, C. NanoCLUST: a species-level analysis of 16S rRNA nanopore sequencing data. *Bioinformatics* **37,** 1600–1601. doi:10.1093/bioinformatics/btaa900. eprint: https://academic.oup.com/bioinformatics/article-pdf/37/11/1600/50361068/btaa900.pdf (Oct. 2020).
64. Rodríguez-Pérez, H., Ciuffreda, L. & Flores, C. NanoRTax, a real-time pipeline for taxonomic and diversity analysis of nanopore 16S rRNA amplicon sequencing data. *Computational and Structural Biotechnology Journal* **20,** 5350–5354. doi:https://doi.org/10.1016/j.csbj.2022.09.024 (2022).
65. Curry, K. D., Wang, Q., *et al.* Emu: species-level microbial community profiling of full-length 16S rRNA Oxford Nanopore sequencing data. *Nature methods* **19,** 845–853 (2022).
66. Andrews, S. *et al. FastQC: a quality control tool for high throughput sequence data* 2010.
67. De Coster, W. & Rademakers, R. NanoPack2: population-scale evaluation of long-read sequencing data. *Bioinformatics* **39,** btad311. doi:10.1093/bioinformatics/btad311. eprint: https://academic.oup.com/bioinformatics/article-pdf/39/5/btad311/50394865/btad311.pdf (May 2023).
68. Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34,** i884–i890 (2018).
69. Ewels, P., Magnusson, M., Lundin, S. & Käller, M. MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* **32,** 3047–3048 (2016).
70. Douglas, G. M., Maffei, V. J., *et al.* PICRUSt2 for prediction of metagenome functions. *Nature biotechnology* **38,** 685–688 (2020).
71. Segata, N., Izard, J., *et al.* Metagenomic biomarker discovery and explanation. *Genome biology* **12,** 1–18 (2011).
72. Dixon, P. VEGAN, a package of R functions for community ecology. *Journal of vegetation science* **14,** 927–930 (2003).
73. Shen, W. & Ren, H. TaxonKit: A practical and efficient NCBI taxonomy toolkit. *Journal of Genetics and Genomics* **48.** Special issue on Microbiome, 844–850. doi:https://doi.org/10.1016/j.jgg.2021.03.006 (2021).
74. Di Tommaso, P., Chatzou, M., *et al.* Nextflow enables reproducible computational workflows. *Nature biotechnology* **35,** 316–319 (2017).
75. *Anaconda Software Distribution* version Vers. 23.9.0. 2024.
76. Kurtzer, G. M., Sochat, V. & Bauer, M. W. Singularity: Scientific containers for mobility of compute. *PloS one* **12,** e0177459 (2017).