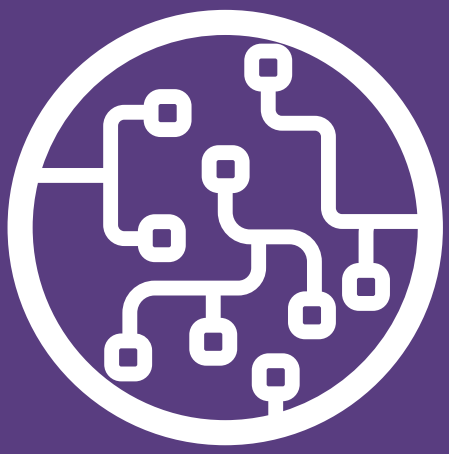
UMAG
Universidad de Magallanes

Titulo por definir



Jacqueline Aldridge Águila

UMAG
Universidad de Magallanes

Titulo por definir

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

Declaración de Autenticidad

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	Titulo por definir
Autor	Jacqueline Aldridge Águila
Grado	Magíster en Bioinformática
Facultad	Facultad de Ingeniería

Fecha Julio, 2024

Agradecimientos

agradecimientos

1

Introducción

Las tecnologías de secuenciación de próxima generación llegaron a cambiar el paradigma de la biología molecular, permitiendo la secuenciación de grandes volumenes de información, lo que antes no era posible. La secuenciación de genomas completos, regiones especificas del genoma, mutaciones, transcriptomas y metagenomas se han vuelto análisis comunes. Sin embargo, a medida que las nuevas tecnologias de secuenciación permitieron la generación de datos que conllevan nuevos análisis ha surguido la necesidad de desarrollar nuevas herramientas bioinformáticas pensadas en estos análisis de datos y que consideren las diferentes caracteristicas de cada tecnología de secuenciación.

Un ejemplo de esto, es la secuenciación del gen 16S la cuál en con el auge de las NGS ampliamente para la caracterización de comunidades microbianas. Con las tecnologías de short-reads, como Illumina, se han desarrollado herramientas bioinformáticas que permiten realizar la asignación taxonómica de secuencias parciales del gen 16S. Con la llegada de las tecnologías de long-reads, como Oxford Nanopore, se ha abierto la posibilidad de secuenciar el gen 16S completo, permitiendo tener una resolución taxonómica mayor. A pesar del aumento de herramientas que permiten procesar datos de secuenciación de tercera generación, la mayoría de estas herramientas están enfocadas en un usuario con conocimientos informáticos o al menos de línea de comando, siendo solo unas pocas herramientas las que permiten realizar análisis de manera más amigable para el usuario.

1.1 | Introducción

1.2 | Objetivos

1.2.1 | Objectivo general

Desarrollar un flujo de trabajo y plataforma automatizada y user-friendly para la asignación taxonómica, caracterización y procesamiento de secuencias del gen 16s secuenciadas por Oxford

Nanopore.

1.2.2 | Objetivos específicos

- Selección y testeo de las mejores herramientas.
- Desarrollar pipeline automatizado que integre clustering de secuencias y post procesamiento de los datos.
- Desarrollar plataforma web de análisis que integre flujo de trabajo automatizado.

1.3 | Descripción del documento

1.4 | Motivación

- La falta de herramientas computacionales para el análisis de datos con Nanopore, lo que conlleva a que muchas personas opten por no utilizar esta tecnologia de secuenciación por no poder llevar a cabo los analisis, sin importar que se pueda tener una resolución taxonomica mayor que con tencologías de lecturas cortas.

2

Marco teórico y estado del arte

Esta tesis busca presentar una nueva alternativa para la caracterización de comunidades microbianas utilizando secuenciación de tercera generación. Para ello se desarrolló un flujo de trabajo automatizado que permite realizar el procesamiento de los datos de secuenciación (control de calidad, asignación taxonómica, análisis de diversidad, predicción funcional y caracterización de grupos).

Debido a que la ejecución de pipelines bioinformaticos requiere conocimiento de línea de comando y contar con recursos computacionales, se desarrolló una aplicación que permite al usuario abstraerse del conocimiento computacional requerido al analizar datos. Una vez que el usuario sube sus datos a la plataforma web, ésta envia los datos, ejecuta el flujo de trabajo y presenta los resultados en forma de gráficos y tablas en la plataforma.

A continación se presenta el marco teórico y estado del arte de los conceptos necesarios para el desarrollo de esta tesis, como qué es la microbiota, el gen 16S rRNA, tecnologías de secuenciación y sus usos, las diferentes herramientas bioinformáticas para el análisis de datos y herramientas para el desarrollo de la aplicación web.

2.1 | Estudio de microbiota a traves del gen 16S y tecnologías de secuenciacion

2.1.1 | 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 [1], animales [2] o plantas [3], y también en ambientes naturales como el océano [4] y el suelo [5]. Estos organismos que componen la microbiota se encuentran en un estado de simbiosis junto con el huesped, contribuyendo en funciones vitales como la homeostasis, regulación del sistema inmune, digestión de alimentos, producción de vitaminas, protección ante enfermedades y agentes patógenos [6-9]. Sin embargo,

una disbiosis o una baja diversidad en la microbiota se puede asociar a una desregulación en el organismo huesped, 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 [8, 9].

La composición de la microbiota va cambiando dependiendo del área de estudio, pudiéndose encontrar diferentes microorganismos en las cavidades orales, zonas intestinales, genitales, cutáneas o tracto respiratorio [10].

En la naturaleza los microorganismos cumplen un rol fundamental en los ciclos bioquímicos del nitrógeno, carbono y fósforo [11, 12], como también en los procesos de desnitrificación, nitrificación y mineralización [11, 12]. 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 [13]) y de las características de éste como la temperatura, hidaratación, profundidad, cantidad de carbono [14]. En el caso de las plantas, se ha demostrado que la microbiota presente ayuda a la adquisición de nutrientes [15], crecimiento, salud y resistencia a enfermedaddes [16-19].

La microbiota humana se puede ver afectada por diferentes factores, como los hábitos alimenticios, estilo de vida, uso de antibióticos, edad, estrés, entre otros [8]. La interacción con el medio ambiente también influye, habiendo estudios que identifican cambios en la microbiota de recién nacidos, infantes y adultos que viven con animales [20-22], como también cambios en la microbiota intestinal y cutánea en niños que interactuan con la naturaleza, plantas o suelo, identificando aumento en las vías inmunoreguladoras en comunidades microbianas cercanas a la naturaleza [23].

Conocer la diversidad microbiana asociada a organismos multicelulares permite ahondar en la relación existente entre microbios y la salud de los seres vivos, asi como conocer microorganismos patógenos que causan enfermedades infecciosas, ayuda al diagnóstico y permite tomar acciones oportunas. Este conocimiento ayuda modular nutrición, salud y enfermedad a través del estudio del microbioma tomando en consideración los distintos factores asociados al estilo de vida.

2.1.2 | ARN Ribosomal 16S

El ARN ribosomal 16S es un gen perteneciente a la subunidad menor 30S que codifica el rRNA bacteriano y se encuentra en todas las bacterías. Esta compuesto por 1542 pares de bases aproximadamente, divididas en 9 regiones hipervariables entrelazadas con regiones constantes [24]. Las regiones constantes son compartidas por todas las bacterias, mientras las regiones variables presentan cierto grado de variabilidad entre las especies.

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

El uso de la macromolécula del ARN ribosomal 16S para el estudio de relaciones filogenéticas y de bacterias fue propuesto por Carl Woese a principios de 1970 [25]. Sus 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 asig-

nación taxónomica, y abordable para análisis bioinformáticos) hacen que hoy en día sea el marcador molecular más utilizado para la identificación de bacterias y comunidades microbianas [26-29].

Las regiones variables permiten llevar a cabo la caracterización de los microorganismos, siendo la metodología más utilizada el secuenciar parcialmente el gen 16S, es decir, secuenciar una o dos regiones y realizar la asignación taxónomica en base a la región secuenciada. 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 secuenciar influye en los resultados de la comunidad y en la diversidad de microorganismos que se caracteriza [30-33].

El gold standard para la identifición de bacterias durante muchos años fue el cultivo convencional en laboratorio, sin embargo, el cultivo puede durar desde días a semanas o incluso meses, y en algunos casos, hay bacterias que no se logran cultivar en laboratorio [34]. Es por esto, que las tecnologías de secuenciación de nueva generación se presentaron como una buena alternativa y se empezaron a usar masivamente para secuenciar el gen 16S y caracterizar comunidades al permitir secuenciar comunidades complejas (y no solo aislados) y al permitir secuenciar millones de lecturas al mismo tiempo [26], haciendo que la forma de caracterizar bacterias sea más estándar y abordable al día de hoy [35, 36].

2.1.3 | Secuenciación de ADN

Todo ser vivo cuenta con una molécula de ADN que contiene la información genética del individuo. Ésta información se códifica a través de bases nitrogenadas: adenina (A),timina (T), guanina (G) y citosina (C) [37]. Para determinar esta secuencia de nucleótidos se han desarrollado diferentes técnicas las que se conocen como secuenciación de ADN.

Las tecnologías de secuenciación se pueden dividir en tres generaciones, cada una con diferentes características, como los largos de las moleculas a secuenciar, porcentaje de error, costo y cantidad de información que secuencian (*throughput*). Las tecnologías de secuenciación de nueva generación (NGS) involucran las tecnologías de segunda y tercera generación (short-reads y long-reads respectivamente) y se diferencian de la primera generación de secuenciación por la cantidad de información que permiten obtener en la secuenciación, y por haber reducido notablemente los costos y tiempos de secuenciación, pero presentando un porcentaje de error mayor [38].

Independiente de la tecnología de secuenciación a utilizar, el proceso de secuenciación de ADN se puede dividir en tres etapas: preparación de librería, secuenciación y análisis de las datos. Durante la preparación de las librerias el ADN se fragmenta en tamaños manejables para el secuenciador que luego son secuenciados. El resultado de la secuenciación es un conjunto de secuencias de nucleótidos conocidas como *read* o lectura. Debido a las diferentes caracterisitcas de cada generación, al analizar las secuencias los análisis bioinformáticos y herramientas a utilizar también cambian dependiendo principalmente de la precisión del secuenciador utilizado y el tamaño de las lecturas producidas [39].

La Figura **??** presenta una comparativa entre las principales tecnologías de secuenciación de ADN, mostrando las diferencias entre las tecnologías de primera, segunda y tercera generación. En las siguientes secciones se detalla el funcionamiento de cada tecnología y sus caracteristicas más importantes.

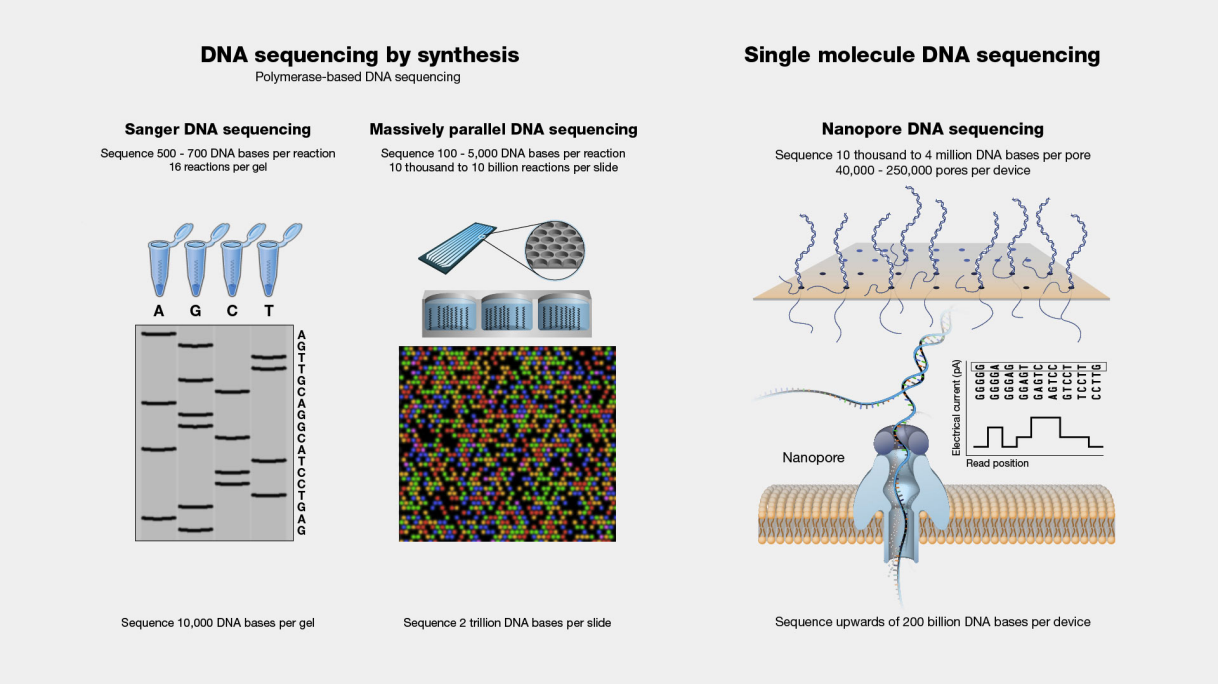


Figura 2.2. Comparativa tecnologías de secuenciación de ADN[1].

Primera generación: Sanger

Frederick Sanger introdujó la primera generación de tecnologías de secuenciación en el año 1977. El método Sanger utiliza el principio de "terminación de cadena" para secuenciar material genético [40]. Durante la PCR se añaden nucleótidos modificados (ddNTPs) marcados con fluoroforo que al ser incorporados en la cadena de ADN detienen la elongación de ésta. Una vez que la reacción ha terminado, se realiza una electroforesis capilar que separa las moleculas de ADN por tamaño y con alta resolución para poder determinar la secuencia de la cadena, generando una única secuencia de como máximo 800 pares de bases por corrida [41].

La principal ventaja de este método es su alta precisión (99.999 %), siendo ampliamente utilizado en diagnóstico clínico para la detección de mutaciones puntuales. Además de no requerir capacidad de cómputo ni conocimientos bioinformáticos para la generación de la secuencia. Entre sus principales desventajas se encuentras que es un método laborioso, costoso e ineficiente al trabajar en proyectos de secuenciación de gran escala como por ejemplo genomas completos o metagenómica debido a la cantidad de información que se puede obtener en la secuenciación, ya que permite obtener una sola secuencia por corrida.

Debido a su alta precisión, al secuenciar el gen 16S con Sanger se permite la caracterización a nivel de especie, pero también se pueden encontrar algunas limitantes, como por ejemplo que esta técnica permite detectar una sola especie a la vez. En caso de tener una muestra aislada, la secuenciación permitirá determinar la especie asociada al aíslado, pero en el caso de que la muestra presente más de una bacteria, la señal que se obtendra será una mezcla de las diferentes bacterias, por lo que será imposible realizar una caracterización. Esto limita su uso en comunidades complejas

[1] Fuente: https://www.genome.gov/genetics-glossary/DNA-Sequencing

o infecciones polibacterianas [42], donde para identificar la comunidad se deben usar tecnologías de secuenciación de nueva generación.

Segunda generación: Secuenciación de lecturas cortas

Las tecnologías de secuenciación de segunda generación más importantes son Illumina, IonTorrent y Roche 454. Éstas utilizan el principio de síntesis de cadena complementaria y una señal asociada al nucleótido incorporado (que puede ser fluorescencia, iones cargados, entre otros). Esta generación de secuenciadores se caracteriza por ser rápida, de bajo costo, y por permitir obtener simultaneamente millones de moleculas de ADN mediante amplificación clonal o *bridge* PCR, generado en paralelo millones de fragmentos cortos de ADN, de entre 35 a 600 pares de bases, con una precisión de 99.9 %.

Las principales metodologías de secuenciación de segunda generación son secuenciación por síntesis, utilizada principalmente por Illumina, y secuenciación por ligación utilizada por Roche 454 [43]. Illumina es la *NGS* más utilizada para la secuenciación del gen 16S. Este método utiliza secuenciación por síntesis: la primera etapa corresponde a la fragmentación de ADN y ligación de adaptadores. Luego los fragmentos de ADN se unen a la superficie de la celda y se amplifican localmente para formar clusters. En cada ciclo de secuenciación se añaden nucleótidos marcados con fluoróforos que son removidos para continuar con el siguiente ciclo. Finalmente, el secuenciador detecta la señal de fluorescencia y registra la secuencia de nucleótidos. Illumina posee diferentes dispositivos de secuenciación que permiten obtener *reads* de 2x150 a 2x300 pares de bases como máximo.

Algunas de las plataformas de secuenciación de segunda generación utilizan secuenciación *pair-end* donde cada fragmento de ADN se secuencia en ambas direcciones, lo que permite obtener fragmentos de mayor tamaño mediante el solapamento de ambas lecturas.

Dentro de las principales ventajas de las tecnologías de lecturas cortas se encuentra su bajo porcentaje de error, bajo costo, y la gran cantidad de herramientas bioinformáticas ya desarrolladas para trabajar con este tipo de datos [44]. Por otro lado, el tamaño de los fragmentos secuenciados es su mayor limitante, haciendo más compleja o imposible la resolución de regiones repetitivas o análisis más complejos como el ensamblaje de genomas.

La caracterización de comunidades microbianas con Illumina se ha convertido en el estándar, debido a su bajo costo,througput y alta precisión. Sin embargo las tecnologías de lecturas cortas permiten secuenciar solo una parte del gen 16S rRNA, debido al tamaño de las moleculas que secuencian (lecturas de entre 200 a 400pb) [45]. Diversos estudios se han realizado para analizar qué región permite obtener la mayor diversidad posible y de manera más precisa [46, 47]. También se ha determinado que existen grupos taxonómicos que se identifican de mejor manera al utilizar cierta regiones hipervariables [48, 49]. De igual forma se ha determinado que al realizar una secuenciación parcial del gen 16S la resolución taxonómica es menor, llegandose a identificar solo hasta el nivel de género [46].

Es por esto, que las tecnologías de tercera generación, 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 un solo *read*, lo que permite una mayor resolución taxonómica a nivel de especie o incluso de cepa [50, 51].

Tercera generación: Secuenciación de lecturas largas

Las tecnologías de secuenciación de tercera generación buscan superar las limitaciones de las tecnologías de secuenciación de lecturas cortas, permitiendo obtener información genómica más completa y detallada. Llegaron a presentarse como una alternativa innovadora y prometedora al poder generar lecturas desde unas pocas kilobases hasta megabases [52]. Junto con su capacidad de secuenciar sin amplificación y en tiempo real. Secuenciación de genomas completos, secuenciación de regiones repetitivas, estudios de metilación y bases modificadas son ahora posible de manera más sencilla gracias a la secuenciación de lecturas largas. Dentro de esta categoría de plataformas se encuentran Pacific Biosciences (PacBio) y Oxford Nanopore Technologies (ONT).

Mientras PacBio realiza secuenciación por síntesis de moleculas largas, Oxford Nanopore secuencia mediante la medición de las variaciones de corriente mientras las moleculas de ADN pasan a través de nanoporos, utilizando la diferencia de potencial para determinar la secuencia de nucleótidos.

Su principal ventaja frente a las tecnologías de primera y segunda generación es la obtención de lecturas largas de mas de 1kb, llegando incluso a obtener lecturas de 1.5Mb en el caso de Oxford Nanopore y 200kb en el caso de PacBio, lo que permite la resolución de regiones repetitivas o complejas, como también la identificación de variantes estructurales e identificación de modificaciones epigeneticas de manera mucho más sencilla que al utilizar *short-reads*. Dentro de sus principales desventajas se encuentra el porcentaje de error generado por estas plataformas, qué mediante mejoras en la química ha ido disminuyendo pasando de cerca de un 15 % [53] a menos 1 % para Oxford Nanopore con su nueva química Q20+[2] [54] y menor a 0.1 % para PacBio [54].

Con la aparición de las tecnologías de secuenciación de segunda y tercera generación [27, 55], la secuenciación del gen 16S rRNA se convirtió en una técnica masiva para la identificación y caracterización de comunidades microbioanas y de patógenos o aislamiento de bacterias clínicas [29].

2.1.4 | Métricas para la evaluación de la diversidad microbiana

Los índices de diversidad permiten cuantificar y describir propiedades generales de las comunidades, como la riqueza, dominancia, y uniformidad. Existen diferentes índices de diversidad, los cuales se pueden dividir en índices de riqueza y de uniformidad, entendiendo la riqueza como el número de especies presentes en una comunidad (sin importar la cantidad de organismos presentes por cada especie) y la uniformidad como la distribución de las abundancias relativas de cada especies.

A continuación se presentan algunos de los índices de diversidad más utilizados en la caracterización de comunidades microbianas.

Índice de Simpson

El índice de Simpson mide dominancia y representa la probabilidad de que al seleccionar dos individuos aleatorios de una muestra, ambos pertenezcan a la misma especie.

Este índice varía entre 0 y 1. Valores cercanos a 0 indican una alta diversidad y una baja dominancia de algunas especies en especifico, es decir, al seleccionar dos individuos al azar, la probabilidad de que pertenezcan a la misma especie es baja. Valores cercanos a 1 indican una baja diversidad y alta

[2] https://nanoporetech.com/accuracy

dominancia, es decir una distribución más homogenea. (alta probablidad de que al seleccionar dos individuos al azar sean de la misma especie).

$$D = \sum_{i=1}^{S} \left(\frac{n_i(n_i-1)}{N(N-1)} \right)$$

donde:

- S es el número total de especies
- n_i es el número de individuos de la especie i
- N es el número total de individuos en la muestra

Este índice suele expresarse mediante su complemento (1-D) o su inverso (1/D), donde valores cercanos a 1 indican mayor diversidad de especies.

Índice de Shannon

Este índice permite cuantificar la variedad de especies, tomando en consideración sus abundancias relativas.

$$H' = -\sum_{i=1}^{S} p_i \ln(p_i)$$

donde:

- S es el número total de especies
- p_i es la proporción de individuos de la especie i respecto al total de individuos

Valores mayores a 3 indican una diversidad muy alta, valores entre 2 y 3 indican que las especies estan en equilibrio y valores inferiores a 2 indican baja diversidad.

Índice de Chao1

El índice de Chao1 es un estimador de ríqueza que permite obtener un estimado del número total de especies (incluyendo aquellas no detectadas debido a un muestreo insuficiente).

$$\hat{S}_{Chao1} = S_{obs} + \frac{F_1^2}{2F_2}$$

donde:

- S_{obs} es el número de especies observadas.
- F_1 es el número de especies observadas que solo se encuentran una vez (singletons).
- F_2 es el número de especies observadas que se encuentran exactamente dos veces (doubletons).

2.2 | Herramientas

El desarrollo de un flujo de trabajo automatizado para la caracterización de comunidades microbionas requiere el uso de diferentes herramientas bioinformáticas para el procesamiento de las secuencias, control de calidad, asignación taxonómica con base de datos, análisis de diversidad, caracterización funcional, etc. Por otro lado, el desarrollo de la aplicación web que estará enlazada al flujo de trabajo, requiere del uso diferentes tecnologías web, librerías y de bases de datos para el almacenamiento de la información, generación de gráficos, y procesamiento de información. A continuación se presentan las principales herramientas a utilizar durante esta tesis.

2.2.1 | Asignación taxónomica

La asignación taxonómica en el contexto de secuenciación del gen 16S, es el proceso donde dada una secuencia de ADN, se busca identificar a qué organismo pertenece. Con la secuenciación del gen 16S rRNA se obtiene un conjunto de lecturas (secuencias de nucleótidos), donde cada lectura pertenece a una bacteria presente en la muestra. Estas lecturas se procesan y se comparan con bases de datos existentes para poder realizar la asignación taxonómica y poder identificarla. Finalmente lo que se obtiene es un perfil de toda la comunidad bacteriana de la muestra, es decir, todas las bacterias presentes que las herramientas bioinformáticas pudieron detectar, junto con su abundancia relativa.

La metodología a utilizar para realizar la asignación va a depender de la tecnología de secuenciación que se haya utilizado.Para las tecnologías de tercera generación lo más común es utilizar la base de datos de RefSeq para buscar la secuencia más parecida en la base de datos mediante alguna herramienta de alineamiento como blast. Algunas metodologías exploran también la corrección de errores de estas secuencias, algoritmos de maximización de expectativas o clustering de las mismas secuencias para aminorizar el porcentaje de error y mejorar la asignación taxonómica.

Algunas de las más utilizadas para trabajar con datos de Oxford Nanopore se presentan a continuación:

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.

Esta herramienta entrega 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 [56] 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). Además, debido a que la herramienta ha dejado de recibir soporte por parte de los desarrolladores, no se han actualizado las versiones de los software ni se han realizado mejoras.

Esta herramienta entrega 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.

NanoRTax

NanoRTax [57] 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.

EMU

EMU [58] 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.

2.2.2 | 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:

FastQC

FastQC[59] 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 (menor al 0.1 %) 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 [60] 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[61] es una herramienta de alto rendimiento diseñada para el procesamiento de archivos de secuenciación con calidad (fastq), permite realizar filtrado de secuencias (por calidad, largo), recortar extremos de baja calidad, recortar adaptadores, eliminar colas polyA, etc.

MultiQC

MultiQC [62] 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.

Seqkit

PICRUSt2

PICRUSt2 [63] es una herramienta para la predicción funcional utilizando secuencias marcadoras de genes. Generalmente se utiliza el gen 16S rRNA para 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) [64] 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 [65] 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 [66] permite la manipulación de información taxónomica de registros de NCBI de una manera 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.

2.2.3 | Lenguajes de programación y Frameworks

Nextflow

Nextflow [67] 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.

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.

SQLalchemy

Librería de Python que permite la comunicación con base de datos no relacionales de manera sencilla, transformando los registros de la base de datos en objetos 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 [39] 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 que se presentó como la continuación de POSTGRES. Permite el uso de tipos de datos complejos realizar consultas tanto relacionales (SQL) y no relacionales (JSON).

2.2.4 | Gestores de paquetes

Conda

Conda [68] 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.

Apptainer

Apptainer (antes llamado Singularity [69]) simplifica la creación y ejecución de contenedores, asegurando el encapsulamiento de los componentes de softwares necesarios para su reproducibilidad y portabilidad.

3

Flujo de trabajo y Aplicación Web

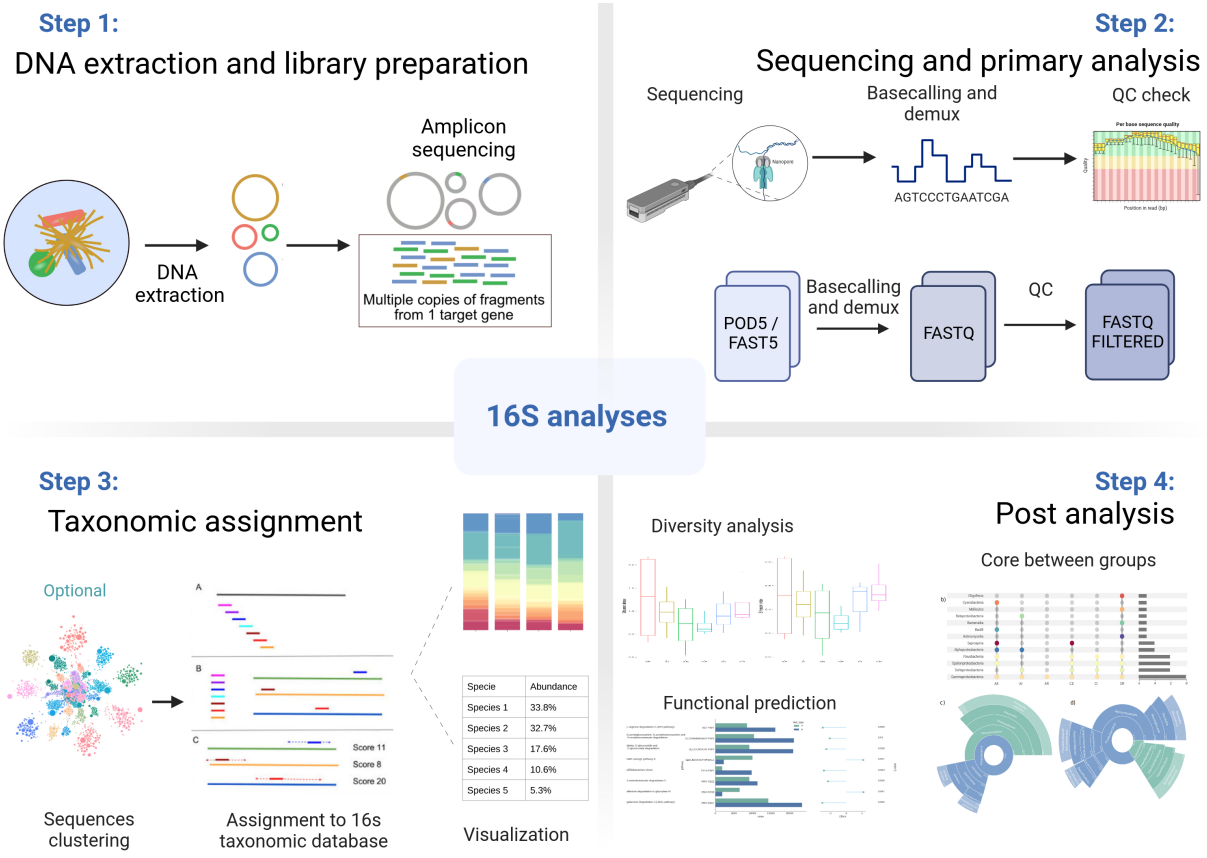


Figura 3.1. Flujo de trabajo estándar para la secuenciación y análisis de secuenciaciones del gen 16S

3.1 | Flujo de trabajo

Se desarrolló un flujo de trabajo automatizado en Nextflow que permite el análisis y la caracterización de secuencias del gen 16S obtenidas mediante dispositivos de secuenciación de Oxford Nanopore. El flujo de trabajo está diseñado de manera modular permitiendo que el usuario pueda personalizar su ejecución, agregando o quitando módulos de análisis mediante un archivo de configuración en formato *YAML*.

El pipeline cuenta con los siguientes módulos:

- Basecalling y demultiplexación (*basecalling*): Este módulo se encarga de realizar el basecalling y demultiplexación de las muestras mediante la herramienta Guppy/dorado.
- Control de calidad (*qc*): Este módulo se encarga de realizar los filtros y control de calidad de las muestras mediante las herramienta Nanoplot y Fastp.
- Asignación taxonómica (*taxonomic_assignament*): Este módulo se encarga de realizar la asignación taxonómica de las muestras mediante la herramienta BLAST y la base de datos de 16S Genbank. Previo a la asignación taxonónimca se realiza un subsampleo de las muestras y se convierten los archivos FASTQ a formato FASTA.
- Indices de diversidad (*diversity*): Este módulo se encarga de calcular los índices de diversidad de Shannon, Simpson y Chao1 mediante el paquete vegan de R.
- Predicción funcional (*functional_predicction*): Este módulo se encarga de realizar la predicción funcional de las muestras mediante la herramienta PICRUSt2 y la búsqueda de vías metabólicas que presenten diferencias significativas entre grupos mediante la herramienta Lefse.

En los módulos de asignación taxonómica, índices de diversidad y predicción funcional mediante scripts en python se resume la información mediantes tablas y se generan gráficos de los resultados obtenidos.

En caso de que el flujo de trabajo se este ejecutando desde la aplicación web, se ejecutará ademas un módulo extra que permite escribir los resultados en la base de datos. Debería tener dos versiones porque hay output no necesarios para cuando es local

3.1.1 | Archivo de configuración

El formato YAML se caracteriza por ser un formato de serialización de datos legible por humanos y fácil de interpretar por máquinas. Tiene una estructura jerarquica que permite la representación de datos de manera más clara y concisa que otros formatos como JSON o XML.

Mediante este archivo de configuración se van a especificar los parámetros necesarios para la ejecución del flujo de trabajo. Además, el archivo de configuración incluye información sobre los archivos de entrada del flujo de trabajo, nombres de las muestras y grupos asociados, todo en un formato de diccionario.

Para especificar parámetros de un módulo se debe indicar el nombre del modulo seguido por *:* y a continuación el nombre del parámetro a modificar y su valor en formato *clave: valor* en la siguente línea. A continuación se detallan los parámetros que se pueden modificar en el archivo de configuración y se presenta un ejemplo de archivo de configuración.

- run: Permite activar o desactivar el módulo. Los valores posibles son “ON” o “OFF”.
- qc.min_length: Permite modificar la longitud mínima requerida de las secuencias. Por defecto es 1000.
- qc.max_length: Permite modificar la longitud máxima permitida de las secuencias. Por defecto es 2000.
- qc.min_qscore: Permite modificar la calidad mínima requerida de las secuencias. Por defecto es 10.
- qc.subsampling: Permite modificar la cantidad de lecturas utilizadas para el subsampleo. Por defecto es 100.000.
- qc.save_reads: Permite activar o desactivar el guardado de los archivo fastq filtrados en el direcotorio de resultados. Valores posibles True o False. Por defecto False.
- taxonomic_assignament.perc_identity: Porcentaje de posiciones identicas en la secuencia de alineamiento.
- taxonomic_assignament.evalue: Valor máximo de evalue aceptado.
- taxonomic_assignament.qcovs: Porcentaje de cobertura mínima de alineamientos con altos puntajes.
- group: Permite ingresar los grupos asociados a las muestras en formato clave:valor.
- input: Permite ingresar el archivo en formato CSV que contiene las columnas samples y fastq.
- outdir: Permite ingresar el directorio donde se guardaran los resultados del flujo de trabajo.
- blast_db: Permite ingresar la ruta al directorio que contiene la base de datos de 16S de Genbank (ver sección **??** para más información).

A continuación se presenta un ejemplo de archivo de configuración donde el módulo de basecalling se encuentra desactivado, y los módulos de control de calidad, asignación taxonómica, índices de diversidad y predicción funcional se encuentran activados. Además se especifican todos los parámetros por defecto del flujo de trabajo, junto con los grupos asociados a las muestras y la ruta al archivo de metadata.

```
basecalling:
  run: 'OFF'
qc:
  max_length: '2000'
  min_length: '1000'
  min_qscore: '15'
  run: 'ON'
  subsampling: '50000'
taxonomic_assignament:
  run: 'ON'
  blast_db: /path/to/db
```

```
diversity:
  run: 'ON'
functional_prediction:
  run: 'ON'
group:
  K1.1: Control K
  L1.1: Control L
  L1.2: Control L
  P1.1: No impactadas
  P1.2: No impactadas
  P7.2: Impactadas
  P8.1: Impactadas
input: samples.csv
outdir: results
```

samples.csv es un archivo de texto separado por comas (formato CSV) que contiene la información de las muestras a analizar mediante las columnas *samples* y *fastq* fastq_2 group. La columna samples debe contener el nombre con el que se quiere identificar las muestra y la ruta al archivo de secuenciación en formato *fastq*.

3.1.2 | Estructura del flujo de trabajo

Basecalling y demultiplexación

Este módulo por defecto se encuentra desactivado y se activa mediante el archivo de configuración. La entrada de este módulo son archivos en formato *POD5* que contienen las secuencias obtenidas en la secuenciación. El basecalling y demultiplexación se realiza mediante la herramienta Guppy/dorado. Por defecto se realizará el basecalling de alta preción (HAC).

Para ejecutar este módulo el usuario deben ingresar los siguientes parámtros en el archivo de configuración:

- *pod5_dir*: Directorio que contiene los archivos de secuenciación en formato POD5.
- *guppy_basecalling_config*: Archivo de configuración a utilizar para hacer el basecalling.
- *guppy_barcoding_kits*: Kit de barcoding a utilizar para hacer la demultiplexación.

Por defecto este módulo entrega un archivo HTML con el reporte de la demultiplexación y el basecalling de las muestras (generado con MultiQC), el cual se almacena en la carpeta QC/multiqc_guppy.html. En caso de que el usuario quiera almacenar los archivos demultiplexados, debe activar la opción *basecalling.save_reads* en el archivo de configuración.

Control de calidad

Este módulo siempre se ejecuta y es el encargado de realizar los filtros y control de calidad de las muestras. La entrada de este módulo son archivos en formato *Fastq*, que pueden ser provenientes del módulo anterior (*basecalling y demultiplexación*) o pueden ser indicados por el usuario a través

del archivo de configuración. El control de calidad se realiza mediante la herramienta NanoPlot (antes y después de los filtros). Los filtros de calidad son llevados a cabo con la herramienta Fastp, utilizando los siguientes parámetros:

- calidad mínima: por defecto 10 o valor ingresado en el archivo de configuración qc.min_qscore (*-q 10*)
- *–cut_mean_quality 10*: Permite recortar bases de baja calidad en los extremos de la secuencia (uso en conjunto con *–cut_front*, *–cut_tail*)
- longitud mínima requerida: por defecto 1000 o valor ingresado en el archivo de configuración qc.min_length (*–length_required 1000*)
- longitud máxima permitida: por defecto 2000 o valor ingresado en el archivo de configuración qc.max_length (*–length_limit 2000*)
- deshabilitar la eliminación de adaptadores (*–disable_adapter_trimming*) (no modificable)
- deshabilitar la eliminación de colas poly g (*–disable_trim_poly_g*) (no modificable)

Además, este módulo cuenta con un script desarrollado en Python que permite graficar la longitud promedio versus la calidad promedio de las muestras después de los filtros.

El output de este módulo es una carpeta llamada *QC* que contiene los siguientes directorios:

- *fastq_filtered*: Directorio con archivos FASTQ de las secuencias filtradas (obtenidos con fastp).
- *fastp_reports*: Directorio con los reportes de los filtros de calidad en formato JSON y HTML (obtenidos con fastp).
- *nanoplot_reports/raw|filtered*: Directorio con los reportes de control de calidad de las muestras antes y después de los filtros (obtenidos con NanoPlot).
- multiqc
- *quality_plot.pdf*: Gráfico de la longitud promedio versus la calidad promedio de las muestras después de los filtros en una ventana de las posiciones de 1400 a 1600 pares de bases.

A continuación se presenta un ejemplo del gráfico de calidad y longitud promedio de las muestras generado por el pipeline (Figura **??**).

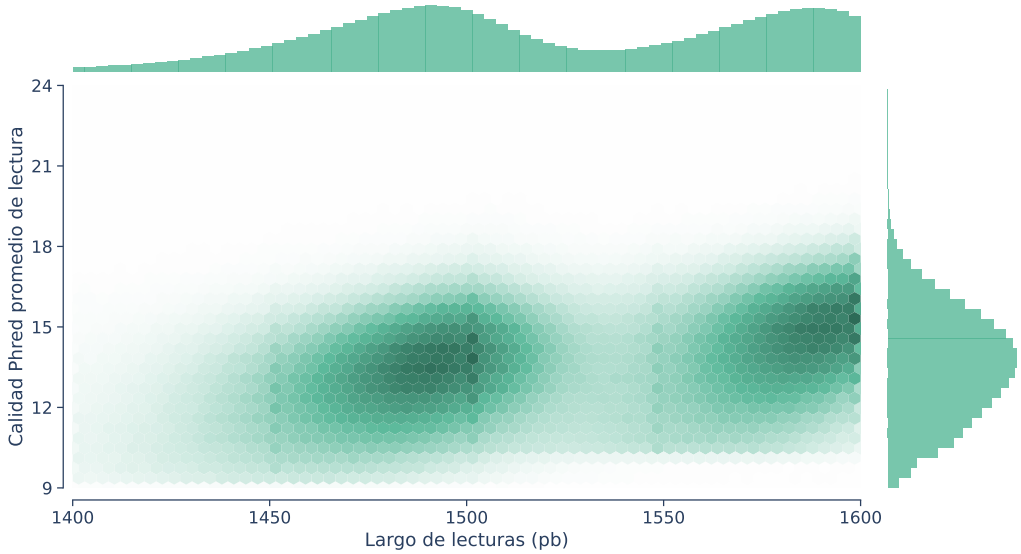


Figura 3.2. Gráfico de calidad vs longitud promedio. En el eje X esta la longitud de las lecturas, mientras en el eje Y se encuentra la calidad promedio.

Asignación taxonómica

Este módulo se ejecuta siempre que el parámetro de asignación taxonómica sea “ON” en el archivo de configuración. La ruta a la base de datos debe ser especificada mediante el parámetro taxonomic_assignament.blast_db. La entrada de este módulo son archivos en formato *FASTQ* que provienen del módulo anterior (*control de calidad*).

Previo a la asignación taxonómica, se realiza un subsampleo de las muestras con la herramienta Seqkit para evitar sesgos en la caracterización de la comunidad microbiana debido a alguna desproporción en la cantidad de lecturas de las muestra. La cantidad de lecturas utilizadas para el subsampling por defecto es 100.000 o el valor ingresado por el usuario en el archivo de configuración *qc.subsampling*. Posterior a ello se convierten los archivos FASTQ a formato FASTA mediante la herramienta Seqkit.

La asignación taxonómica se realiza mediante la herramienta BLAST con la base de datos de 16S de Genbank. Para aceptar un match con la base de datos se requiere un valor de identidad mayor al 97 %, una cobertura mayor al 85 % y un evalue menor a 1e-6.

Posterior a la asignación taxonómica, mediante la herramienta TaxonKit se obtiene el linaje completo de todas las especies asignadas con Blast, este resultado se formatea mediante la herramienta CSVTk. Esto se realiza con el objetivo de poder graficar todas las categorías taxonómicas tanto en los gráficos de barras apilas como en el gráfico circular jerárquico.

Mediante un script en python se realiza un resumen de la asignación taxonómica obtenida con BLAST de todas las muestras y los linajes asociados a cada especie en un solo archivo. Además, se generan archivos por cantidad de lecturas y porcentaje, por cada categoría taxonómica y por muestra y grupo (en caso de ingresarse). Estos archivos son utilizandos como entradas para los scripts que generan los gráficos de barras apiladas y el gráfico circular jerárquico.

En el caso del gráfico circular, se buscan todas las taxonomías compartidas entre las muestras que

tengan una abundance mayor al 1 % en al menos una muestra. Una vez que se identifican, se suman las lecturas asignadas a cada taxonomía en todas las muestra del grupo y se grafica el porcentaje de lecturas asignadas a cada taxonomía en todas las muestras del grupo como un valor único.

El output de este módulo es una carpeta llamada *taxonomic_assignament* que contiene los siguientes directorios:

- blast_out
- plots: Este directorio contiene los siguientes directorios:
  - taxonomy_plots: Directorio con los gráficos de barras apiladas de las taxonomías. Contiene cuatro tipos de graficos:
    - barras apiladas por muestra utilizando el valor porcentual.
    - barras apiladas por muestra utilizando la cantidad de lecturas.
    - barras apiladas por grupos utilizando el valor porcentual.
    - barras apiladas por grupos utilizando la cantidad de lecturas.
  - core_plot: Directorio con los gráficos circulares jerárquicos de las taxonomías compartidas entre las muestras utilizando el número de lecturas. Habrá un gráfico general que busca las similitudes en todas las muestras, y un gráfico por cada grupo ingresado (en caso de ingresar grupos)(Figura **??**).

A continuación se presentan ejemplos de los gráficos generados en este módulo. La figura **??** presenta un gráfico de barras apiladas por muestra utilizando el valor porcentual y la categoría de clase, mientras que la figura **??** muestra un gráfico de barras apiladas por grupos utilizando el valor porcentual.

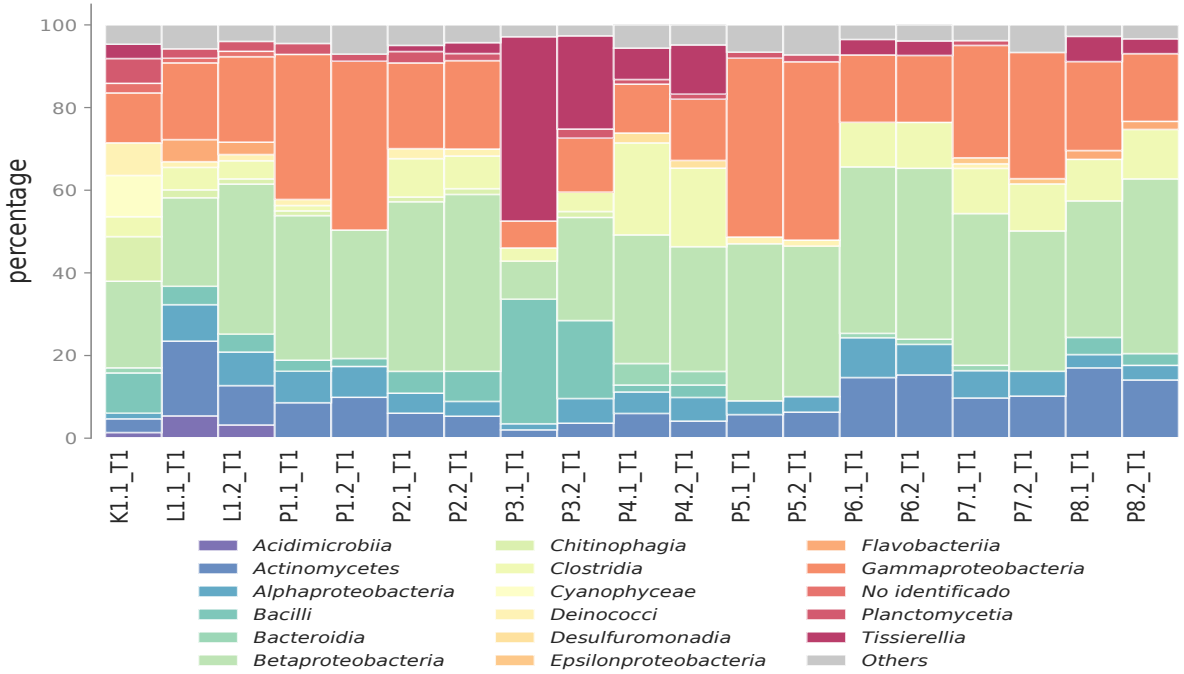


Figura 3.3. Gráfico de barras apiladas por muestra utilizando el valor porcentual y la categoría de clase

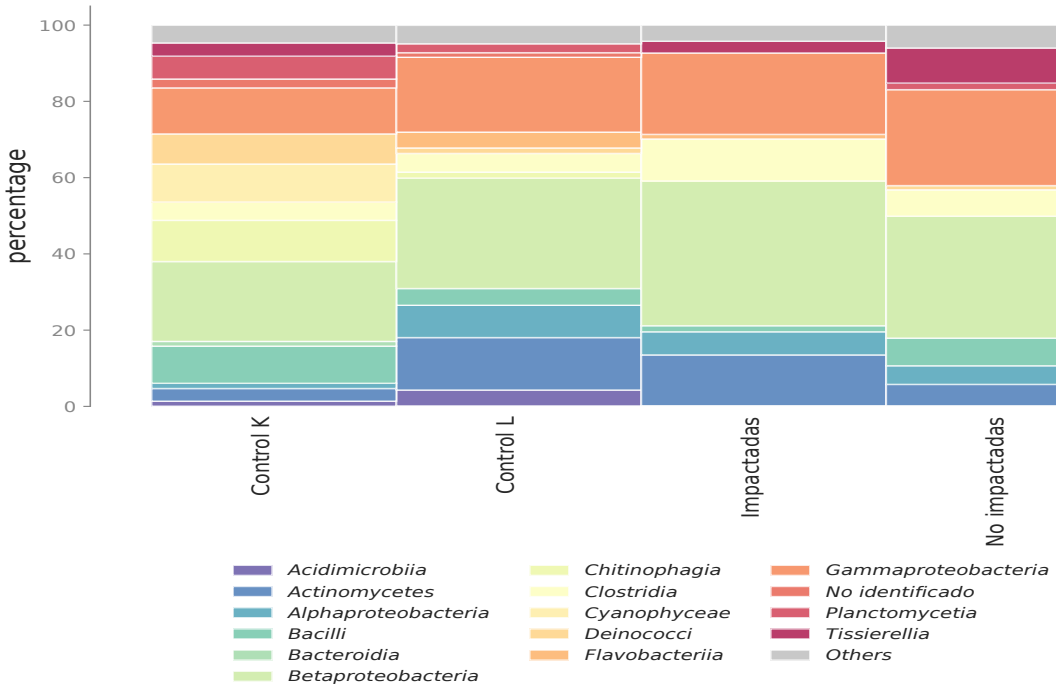


Figura 3.4. Gráfico de barras apiladas por grupos utilizando el valor porcentual y la categoría de clase

La figura **??** presenta un gráfico circular jerárquico generado para el grupo de muestras categorizadas como No impactadas.

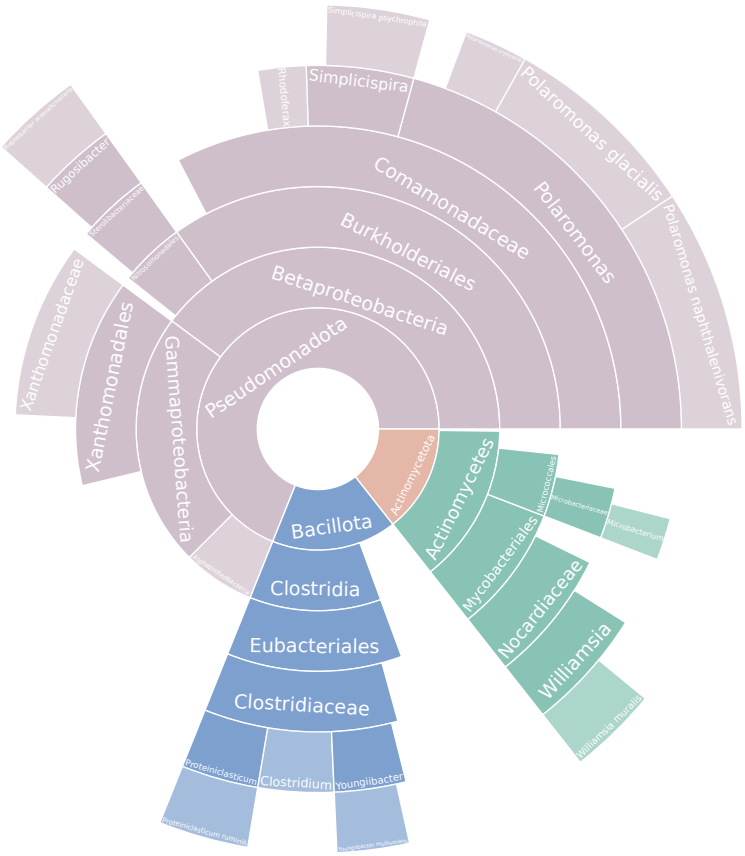


Figura 3.5. Gráfico jerárquico generado para el grupo de muestras categorizadas como No impactadas

Indices de diversidad

Este módulo solo se ejecutara si el parámetro *diversity.run* es “ON” en el archivo de configuración. Para poder ejecutarlo se requiere que el usuario haya ingreado grupos asociados a las muestras en el archivo de metadata.

El cálculo de los índices de diversidad se realizará utilizando el paquete *vegan* de R. La entrada de este módulo es el archivo resumen de blastn obtenido en el módulo anterior (*merge_blast_out*) que contiene en las columnas las muestras y en las filas las especies, y en la intersección la cantiadad de lecturas asociadas. Además del archivo resumen de blast, este módulo requiere los grupos asociados a las muestras en formato diccionario (proveniente del archivo de configuración).

Mediante la función *diversity* se calculan los índices de Simpson y Shannon, y mediante la función *estimateR* se calcula el índice de Chao1. Utilizando la librería ggplot se representa la información de los índices de diversidad en un gráfico de cajas y bigotes. Para el calculo de los indices solo se

considerarán aquellos grupos con al menos 3 muestras.

El output de este módulo es una carpeta llamada *diversity* que contiene un archivo csv con los valores de los índices de diversidad calculados para cada muestra y grupo ingresado (*diversity_index.csv*), junto con el gráfico de cajas y bigotes generado *diversity_boxplot.pdf*

A continuación se presenta un ejemplo del gráfico de cajas y bigotes generado por el pipeline (Figura **??**).

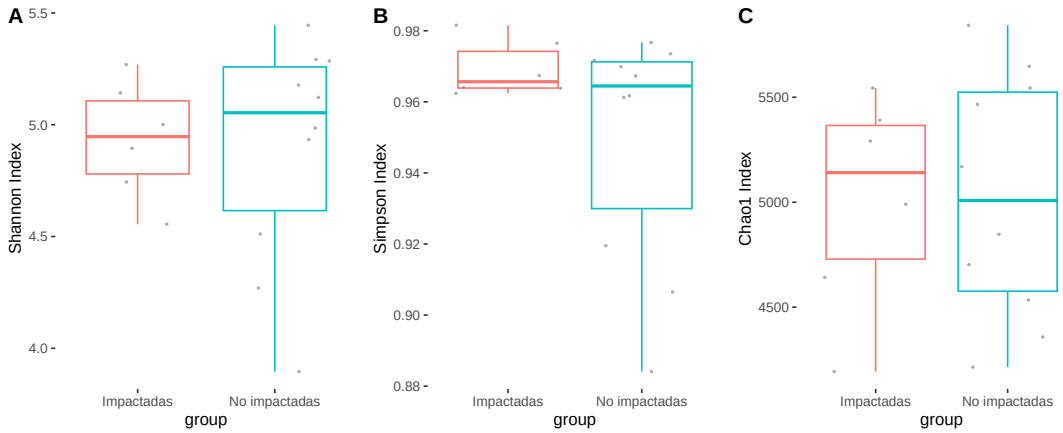


Figura 3.6. Gráfico de cajas y bigotes de los índices de diversidad calculados para cada grupo ingresado

Predicción funcional

Este módulo solo se ejecutara si el parámetro *functional_prediction.run* es "ON" en el archivo de configuración. Para poder ejecutar la diferenciación entre grupos se requiere que el usuario haya ingreado esta información en el archivo de metadata.

La predicción funcional se realizará mediante la herramienta PICRUSt2. Para disminuir los tiempos de ejecucción, se realizará la predicción por muestra y luego mediante un script en python se unirán los resultados en un solo archivo. Para la búsqueda de vías metabólicas que presenten diferencias significativas entre grupos se utilizará la herramienta Lefse (valor de normalización).

El output de este módulo es una carpeta llamada *functional_prediction* que contiene los siguientes directorios:

- *picrust2_out*: Directorio con los resultados de la predicción funcional por muestra. Contiene los siguientes directorios:
  - *EC_metagenome_out*: Directorio con los resultados de la predicción de las enzimas EC.
  - *KO_metagenome_out*: Directorio con los resultados de la predicción de los genes KO.
  - *Pathways_out*: Directorio con los resultados de la predicción de las vías metabólicas.
- *KO.csv*: Archivo resumen con los resultados de la predicción de los genes KO para todas las muestras.
- *EC.csv*: Archivo resumen con los resultados de la predicción de las enzimas EC para todas las muestras.

- *Pathways.csv*: Archivo resumen con los resultados de la predicción de las vías metabólicas para todas las muestras.
- Gráfico de barras con vías metabólicas que presentan diferencias significativas entre grupos

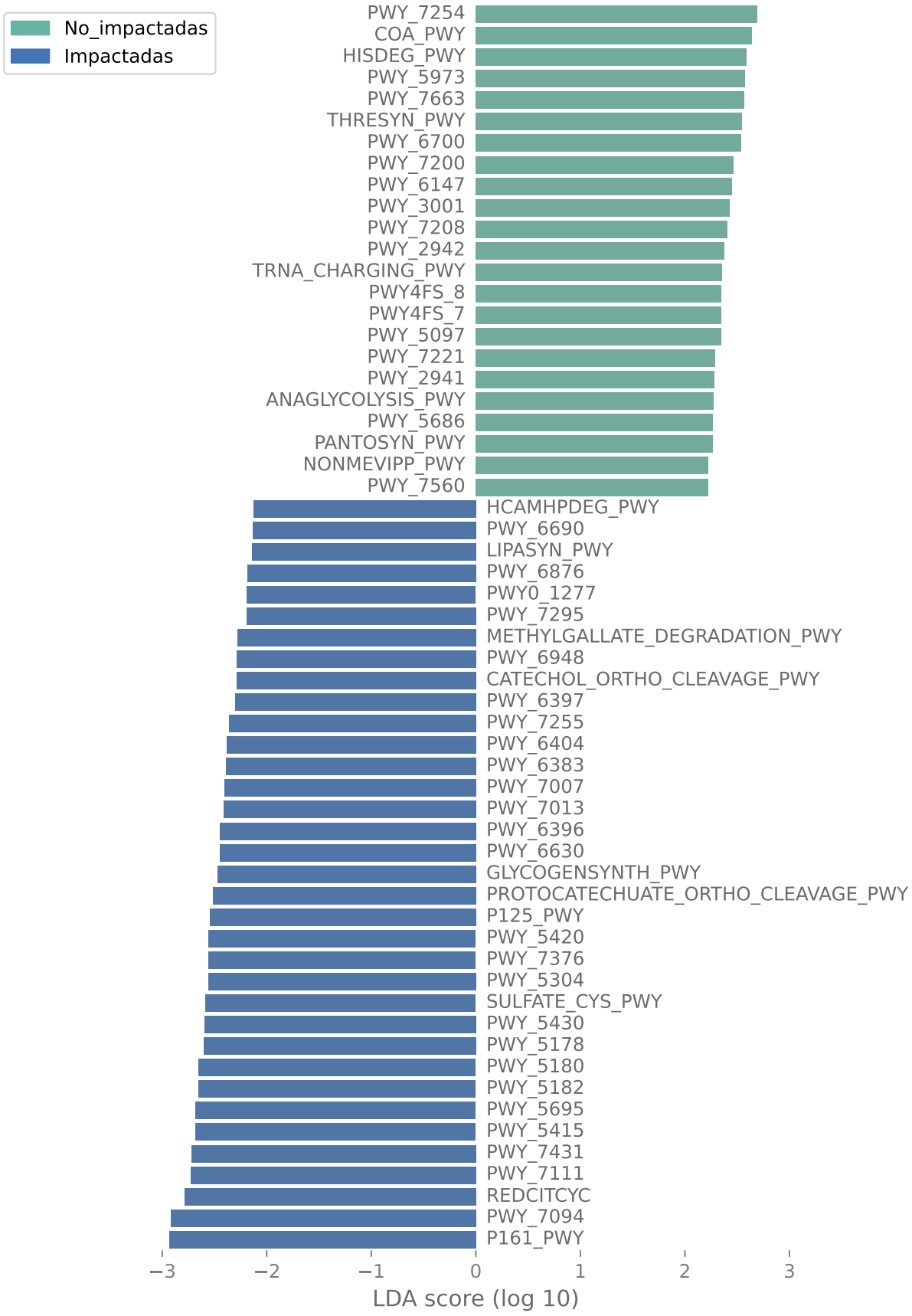


Figura 3.7. Barra de navegación

Ejecución del flujo de trabajo

Para ejecutar el flujo de trabajo se debe descargar el ejecutable de Nextflow desde su página oficial[1]. Además se requiere tener instalado conda para gestionar la instalación de paquetes y herramientas.

Para ejecutar el flujo de trabajo se debe ingresar el siguiente comando en la terminal:

```
nextflow run nanotax-pipeline -profile conda -params-file params.yml
```

La descarga de la base de datos de 16S de Genbank se puede realizar mediante el FTP de NCBI[2].

Se recomienda usar la opción *-resume* para reanudar la ejecución en caso de que se haya interrumpido.

Escritura en la base de datos

archivo que hace params

Desarrollo de contenedores o conda enviroments Para la reproducibilidad del flujo de trabajo se cuenta con dos executers: Conda y Singularity .

[1] https://www.nextflow.io/
[2] ftp://ftp.ncbi.nlm.nih.gov/blast/db/16S_ribosomal_RNA.tar.gz

3.2 | Aplicación Web

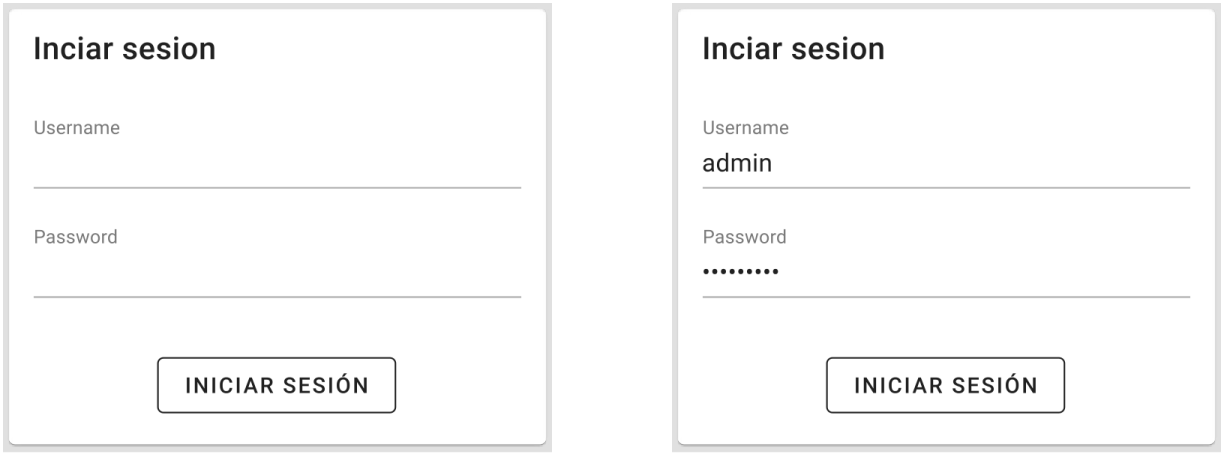
Se desarrollo una aplicación web mediante Vue3 y FastAPI que permite al usuario subir sus archivos de secuenciación y metadata. Con esto el usuario puede abstraerse de tener conocimiento en línea de comando o ejecución de heramientas bioinformaticas y/o flujos de trabajo, ya que mediante la interfaz web el usuario solo debe seleccionar los análisis que desea realizar. La información ingresada por el usuario es guardada en la base de datos y mediante un script se generan los parámetros necesarios para la ejecución del flujo de trabajo. Una vez el pipeline termina de ejecutarse se escriben los resultados en la base de datos. La plataforma lee esta información directa de la base de datos y despliega los resultados en forma de tablas y gráficos en la sección de análisis.

script que genera los parametros de ejecución del flujo de trabajo

A continuación se detalla cada una de las vistas de la aplicación web y su funcionalidad.

3.2.1 | 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 **??**). En caso de que los datos sean incorrectos se mostrará un mensaje de error *"Usuario o contraseña incorrectos"* y deberá ingresar sus credenciales nuevamente.



(a) Vista de inicio de sesión por defecto

(b) Vista de inicio de sesión con mensaje de error al ingresar credenciales inválidas

Figura 3.8. Vista de Inicio de sesión

Registrarse - vista

3.2.2 | Navbar

Una vez que el usuario inicia sesión, va a poder visualizar la barra de navegación de la plataforma en la parte superior de la página.

Figura 3.9. Barra de navegación

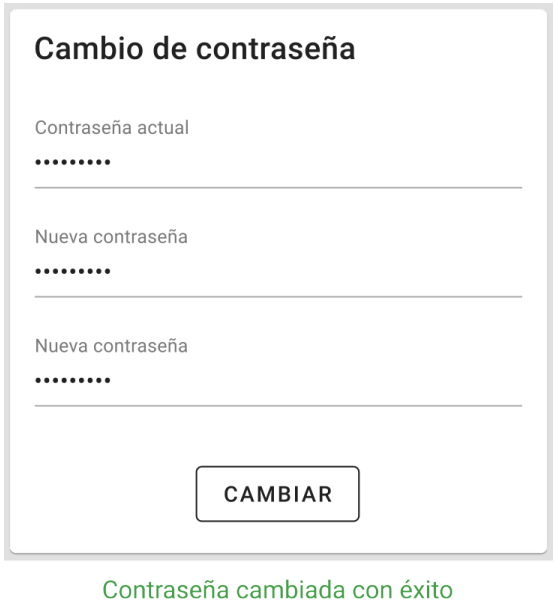
A continuación se describen las funcionalidades de cada una de las secciones de la barra de navegación:

3.2.3 | Cambio de contraseña

Una vez que el usuario validó sus credenciales va a poder acceder a la vista de cambio de contraseña a través de la barra de navegación (Figura **??**).

Para realizar el cambio de contraseña, deberá ingresar su contraseña actual y su nueva contraseña dos veces. En caso de que la contraseña sea cambiada con éxito se mostrará el mensaje *"Contraseña cambiada con exito"* (Figura **??**).

(a) Vista de cambio de contraseña por defecto



(b) Vista de cambio de contraseña realizado con exito

En caso de que la contraseña actual sea incorrecta se mostrará el mensaje de error *"Contraseña incorrecta"* (Figura **??**). En caso de que las contraseñas nuevas no coincidan se mostrará el mensaje de error *"Las contraseñas no coinciden"* (Figura **??**)

(a) Vista de cambio de contraseña al ingresar contraseña incorrecta

(b) Vista de cambio de contraseña al ingresar contraseñas que no coinciden

En el caso de que la nueva contraseña no cumpla los críterios de seguridad (longitud mínima de 8 caracteres y al menos un número) se mostrará el mensaje de error *"La contraseña debe tener al menos 8 caracteres / La contraseña debe tener al menos un número"* (Figura **??**).

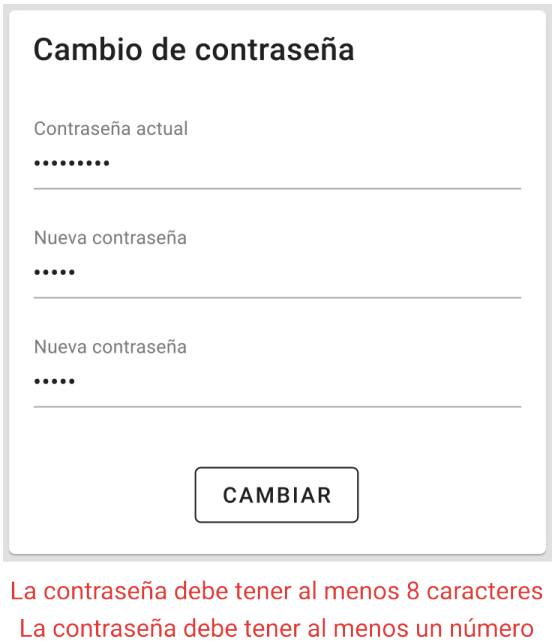


Figura 3.12. Vista de cambio de contraseña al ingresar una nueva contraseña que no cumple con los críterios de seguridad

3.2.4 | Nuevo análisis

En esta sección el usuario deberá ingresar la información del proyecto, datos de secuenciación, y metadata para poder realizar los análisis. El usuario deberá rellenar la información básica del proyecto como, nombre, descripción, tipo de archivos y mediante un archivo en formato (XLXS) deberá ingresar la información de las muestras. Los datos de secuenciación se debe subir a algún directorio del drive del usuario y se debe dar acceso a la cuenta *nanotax.catg@gmail.com*. La figura **??** presenta la vista inicial de la sección de Nuevo análisis.

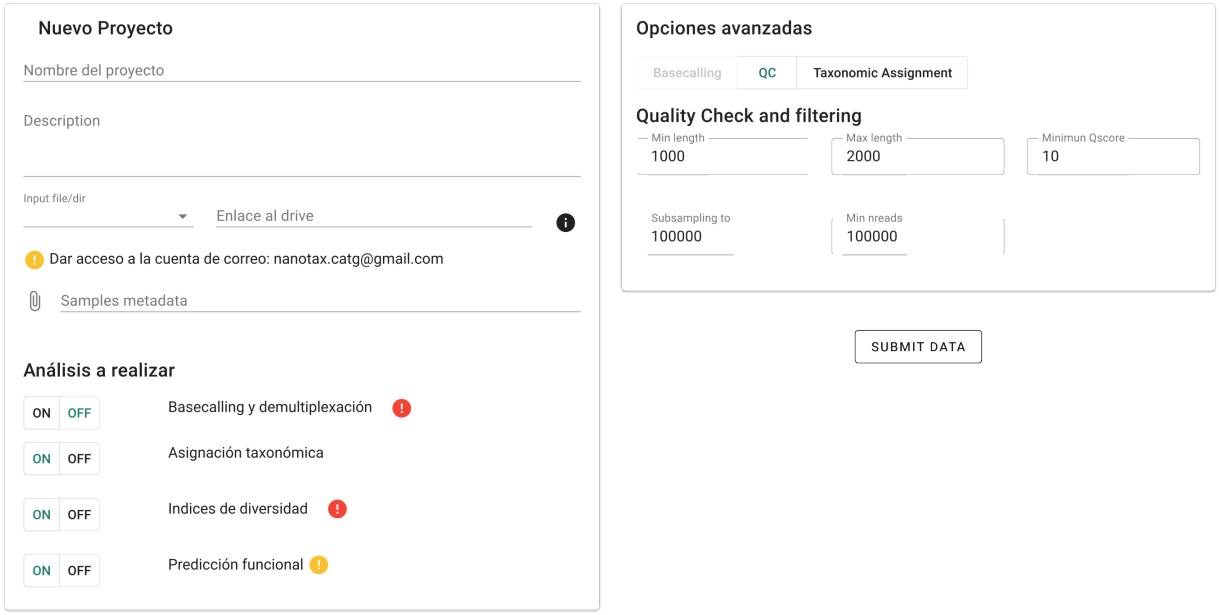


Figura 3.13. Vista por defecto de Nuevo análisis

A continuación se describen los datos que el usuario debe rellenar:

- Nombre: Nombre del proyecto a utilizar en la plataforma (sección de visualización de proyectos y resultados).
- Descripción (opcional): Descripción del proyecto, campo opcional.
- Tipo de archivo a subir (POD5, FASTQ): Archivos de secuenciación que se procesarán:
  - POD5: En caso de querer comenzar desde el proceso de basecalling y demultiplexación de las muestras.
  - FASTQ: En caso de querer saltarse el paso de basecalling y demultiplexación e iniciar directamente con el control de calidad y asignación taxonómica.
- Archivo de metadata en formato XLXS con las siguientes columnas para cada muestra:
  - file: nombre del archivo subido al drive (obligatorio)
  - sample: identificador de la muestra (obligatorio)
  - barcode (opcional): barcode que identifica la muestra (en caso de querer realizar basecalling y demultiplexación)

- group (opcional): 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)

- Análisis a realizar:
  - Basecalling y demultiplexacion
  - Asignación taxonomica
  - Indices de diversidad
  - Predicción funcional

En el lado derecho de la vista se puede visualizar una sección de opciones avanzadas, donde el usuario puede modificar los parámetros por defecto en caso de que quiera modificar el comportamiento del pipeline (gigura **??**). Esta información es seleccionados desde la base de datos la cual almacena los parámetros por defecto del flujo de trabajo.

Cabe destacar que en caso de que el directorio del drive no contenga la información necesaria, el proyecto se subirá correctamente y luego pasara a un estado de datos inválidos Los filtros y control de calidad se realizan siempre por lo que no aparecerá la opción en la lista de análisis. Por defecto basecalling y demuliplexación se encuentra deshabilitado, en caso de que el usuario desee realizar este análisis deberá seleccionarlo, y al hacerlo se desbloqueará la sección de configuración de este análisis (Figura **??**).

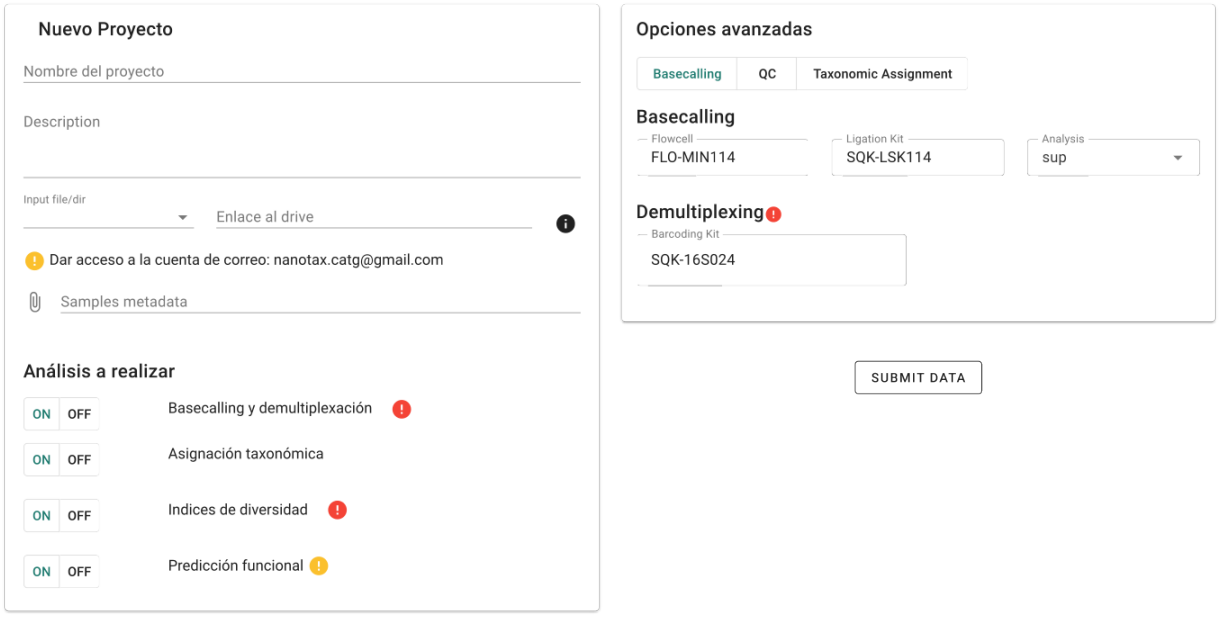
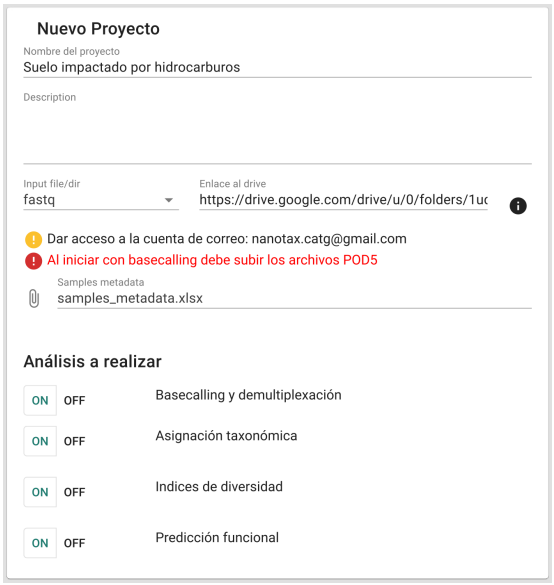


Figura 3.14. Vista de Nuevo análisis habilitando la opción de basecalling y demultiplexación

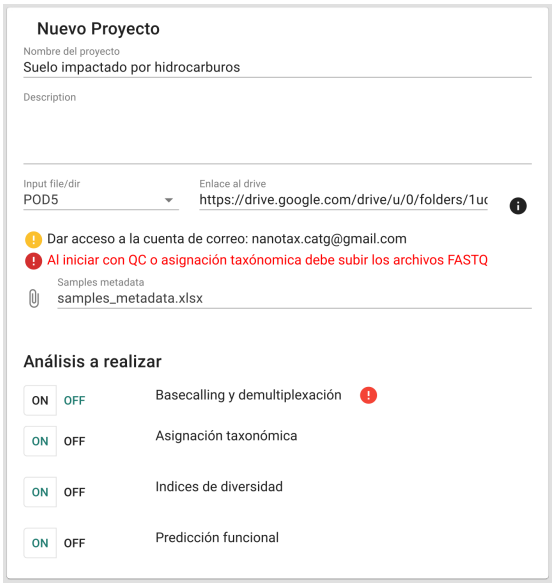
Una vez que el usuario presione al botón *Subir proyecto*, la plataforma realiza un proceso de validación para verificar que toda la información subida por el usuario sea correcta. En caso de no serla,

la plataforma no permitirá subir el proyecto y podrá presentar alguno de los siguientes mensajes de error:

- En caso de no completar el nombre del proyecto o el enlace al directorio del drive ambos campos pasarán a estar en color rojo (figura **??**).
- En caso de no seleccionar el tipo de archivo a subir, este campo pasará a estar en color rojo y se presentará el siguiente mensaje: *Debe seleccionar el formato de los archivos de entrada* (figura **??**).
- En caso de seleccionar el formato de archivo *POD5* y no haber seleccionado el proceso de basecalling y demultiplexación como inicio se presentará el mensaje: *Al iniciar con basecalling debe subir los archivos POD5* (figura **??**).
- En caso de seleccionar el formato de archivo *FASTQ* y haber seleccionado el proceso de basecalling y demultiplexación como inicio se presentará el mensaje: *Al iniciar con QC o asignación taxonómica debe subir los archivos FASTQ* (figura **??**).
- En caso de que el archivo de metadata no cuente con todas las columnas necesarias se pueden presentar los siguientes mensajes de errores (figura **??**) no todos, solo lo que falte:
  - *El archivo de metadata le falta la columna file*
  - *El archivo de metadata le falta la columna sample*
  - *El archivo de metadata le falta la columna barcode*: Solo en caso de seleccionar basecalling y demultiplexación como inicio del pipeline.
  - *El archivo de metadata le falta la columna group*: Solo en caso de querer realizar análisis por grupos (índices de diversidad).

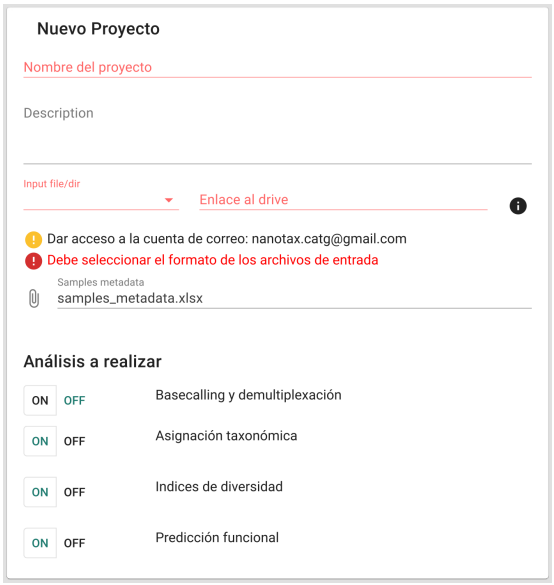


(a) Vista de nuevo análisis: Error al seleccionar el tipo del archivo

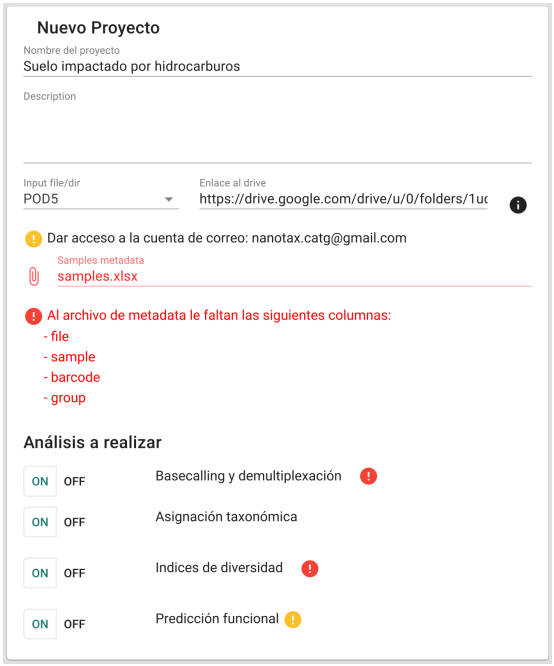


(b) Vista de nuevo análisis: Error al seleccionar el tipo del archivo

Figura 3.15. Vista de nuevo análisis: Error al seleccionar el tipo del archivo



(a) Vista de nuevo análisis: Errores por falta de información



(b) Vista de nuevo análisis: Errores en el archivo de metadata

Figura 3.16. Vista de nuevo análisis: Errores

3.2.5 | Resultados/Proyectos

hacer 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.

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 a la plataforma de computo. 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.

3.2.6 | Resultados de un proyecto en especifico

Una vez que el pipeline haya finalizado su ejecución, la plataforma permitirá al usuario acceder a los resultados de cada proyecto leyendo los resultados desde la base de datos y desplegando la

información en la sección de resultados de cada proyecto. 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, solo se mostrarán las secciones indicadas por el usuario.

3.2.7 | Información básica de las muestras

Esta sección se desplegará siempre en la plataforma y cuenta en el lado izquierdo con una tabla con información básica de las muestras y en el lado derecho un gráfico que representa la calidad y tamaño promedio de las lecturas. La tabla esta compuesta por los siguientes elementos:

- Nombre de la muestra: Nombre indicado en el archivo de metadata al ingresar el proyecto.
- Grupo: Grupo asociado a la muestra en el archivo de metadata (en caso de ingresar grupo).
- Total de lecturas: Cantidad de lecturas previo a los filtros de calidad.
- Calidad promedio: Calidad promedio en formato phred antes/despues de los filtros de calidad.
- Largo promedio: Largo promdio antes/despues de los filtros de calidad
- Lecturas después de los filtros: Cantidad de lecturas luego de los filtros de calidad.
- Nota: Si la muestra fue descartada por no contar con la cantidad suficiente de lecturas se informará 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, este valor por defecto es 100.000, pudiendo ser modificado por el usuario en las opciones avanzadas al ingresar el proyecto.

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

En la parte inferior del gráfico hay una nota que indica en que rangos de tamaño se encuentran la mayoria de las lecturasMuy generico?.

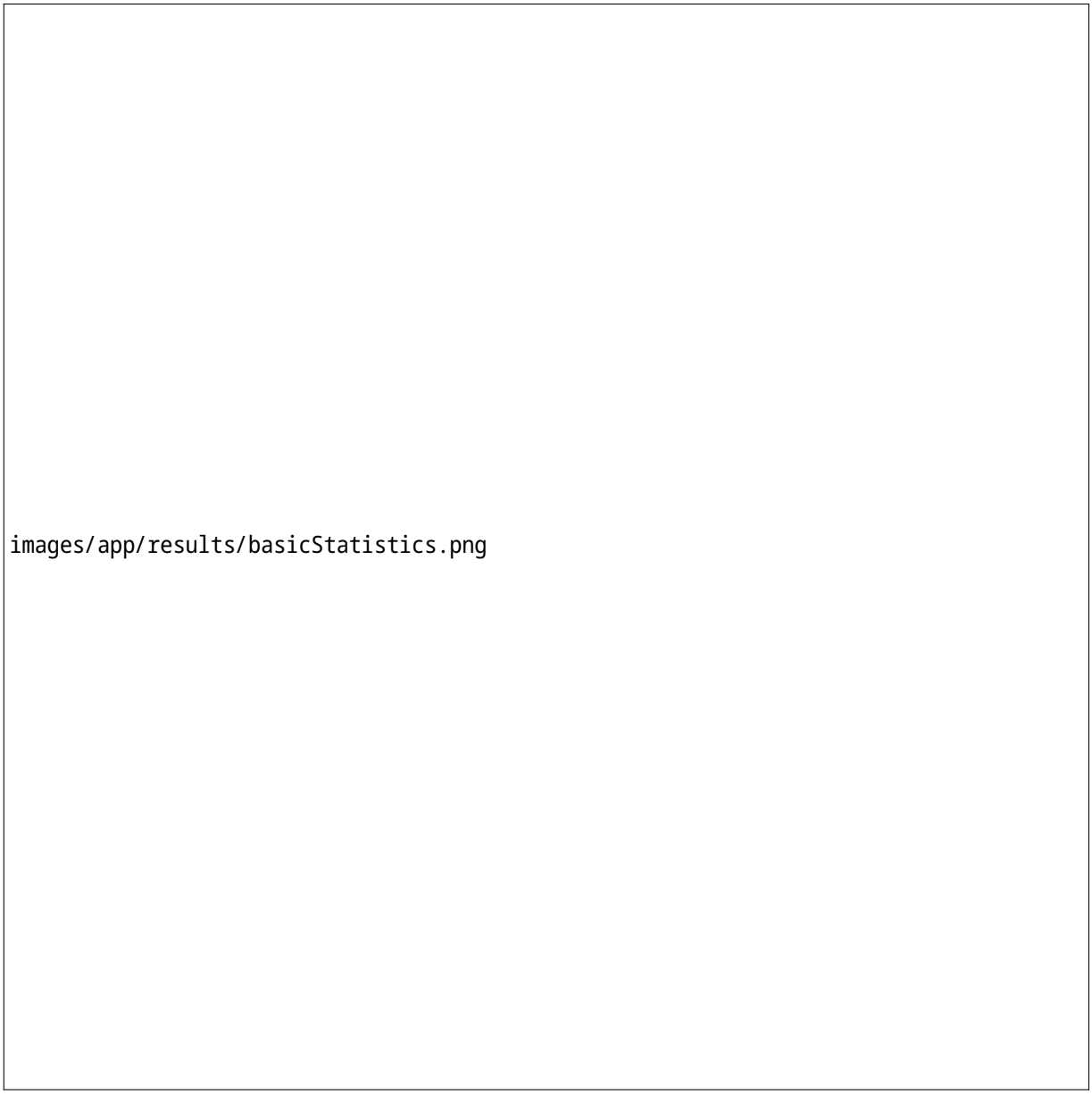


Figura 3.17. Estadisticas básicas (resultados)

3.2.8 | Asignación taxonómica

En la parte superior de esta sección se pueden visualizar pestañas que representan cada categoría taxonómica (especie, género, familia, orden, clase y filo), las cuales permiten ajustar la información presentada en esta sección (gráfico de barras apiladas y tabla). Por defecto se presenta la información para la categoría de especie.

Debajo de las pestañas en el lado izquierdo hay un gráfico de barras apiladas que permite visualizar la abundancia de las taxonomías en cada muestra. El usuario puede interactuar con el gráfico modificando la visualización a través de los botones que se encuentran en la parte inferior, pudiendo visualizar la información en porcentaje o en cantidad de lecturas, como también pudiendo modificar el porcentaje mínimo para crear la categoria *Otros*. En caso de que el usuario hubiera ingresado

información de grupos asociados a las muestras, se podrá visualizar un nuevo grupo de botones que permite al usuario visualizar la información de las taxonomías por grupo o por muestra. La leyenda del gráfico de barras apiladas presenta solo las 10 taxonomías con mayor abundancia. Por defecto, todas aquellas taxonomías que tienen un porcentaje menor a un 1 % son agrupadas en una nueva taxonomia llamada *Otros*.

En el lado derecho, hay una tabla que permite visualizar el detalle de la información presentada en el gráfico. Se puede visualizar además un campo de texto que permite al usuario buscar una taxonomía en especifico y visualizar su abundancia o cantidad de lecturas en todas las muestras.

Ambos componentes, la tabla y el gráfico de barras apiladas se ajustan automáticamente para presentar la información requerida por el usuario, es decir, cada vez que el usuario selecciona una nueva categoría taxonómica en las pestañas, se vuelve a generar la información proporcionando una visualización clara y detallada de los datos.

Figura 3.18. Estadisticas básicas (resultados)

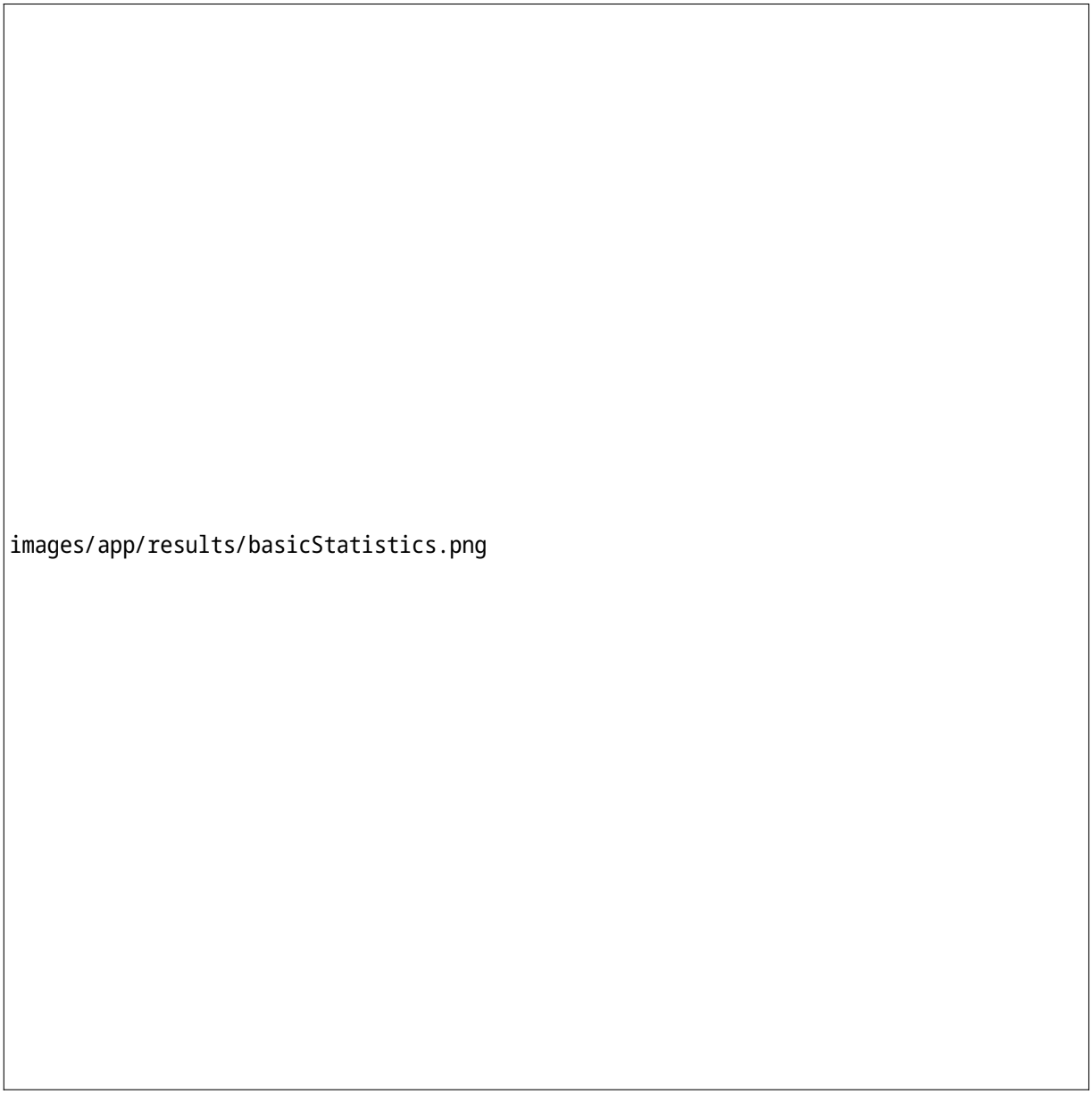


Figura 3.19. Estadisticas básicas (resultados)

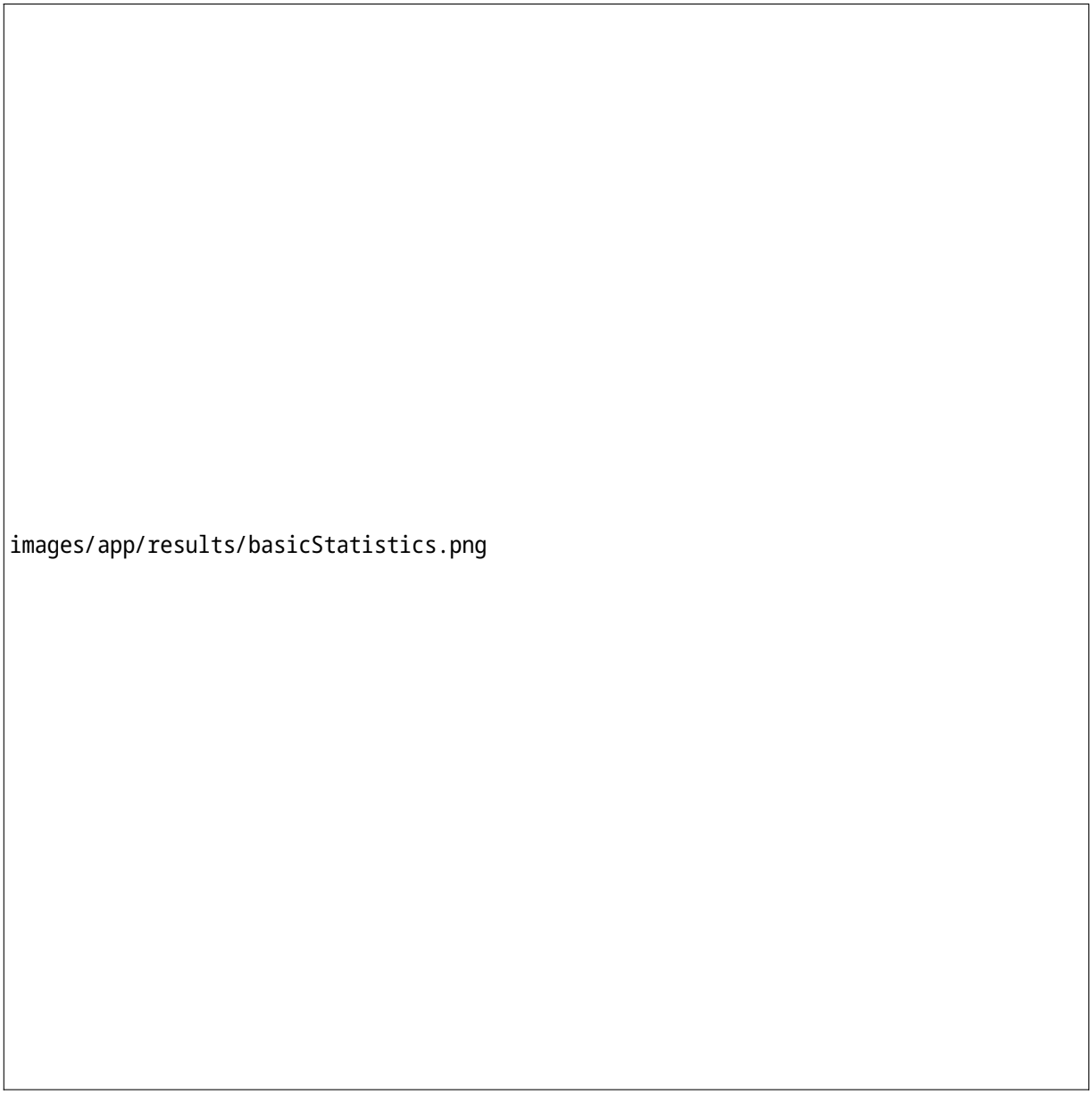


Figura 3.20. Estadisticas básicas (resultados)

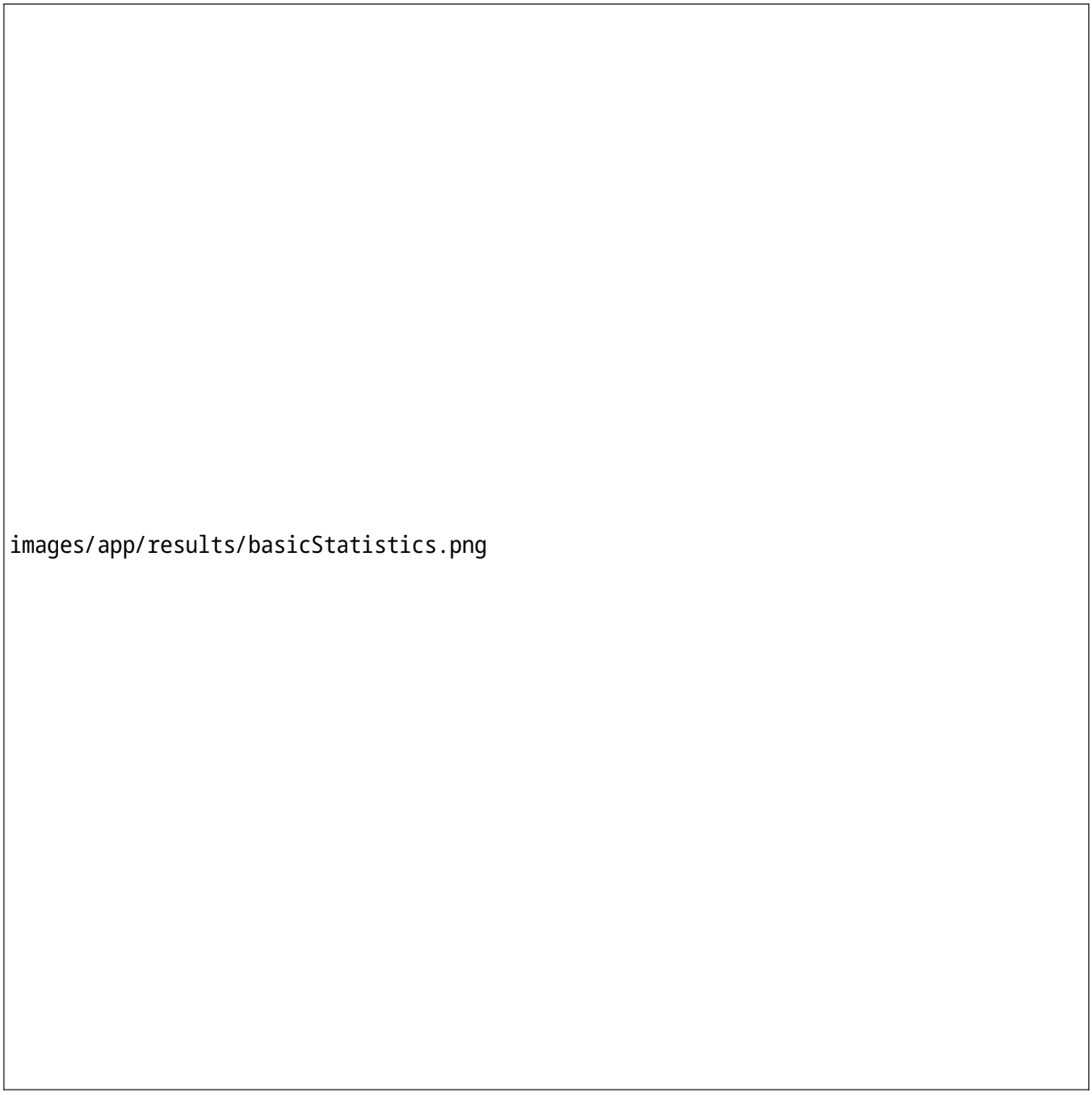


Figura 3.21. Estadisticas básicas (resultados)

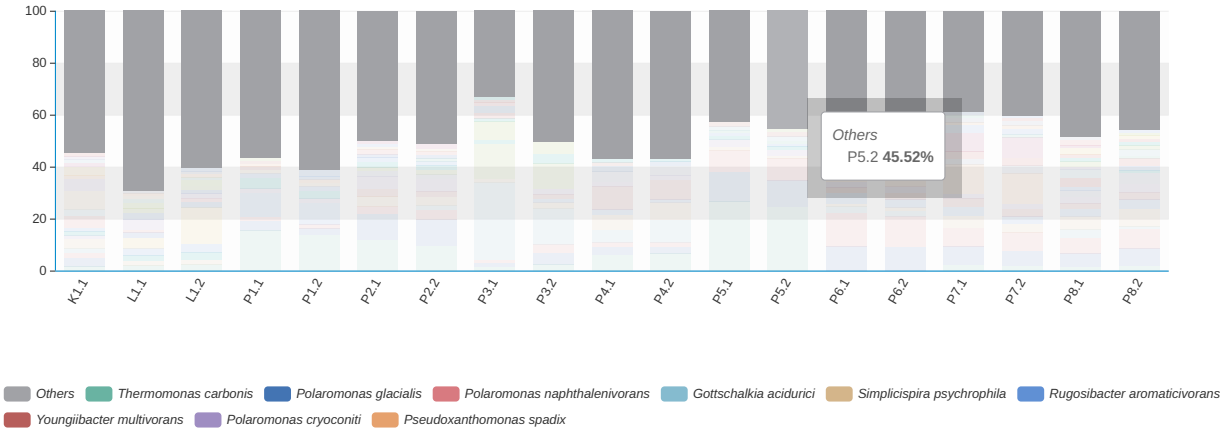


Figura 3.22. Estadisticas básicas (resultados)

3.2.9 | Similitud entre las muestras

la info es la suma de todo?

Esta sección representa las taxonomías compartidas por las muestras mediante un gráfico circular de anillos jerarquicos. Cada nivel del gráfico representa una categoría taxonómica, siendo la más interna especie, y la más externa filo. Mientras más grande el diamentro del anillo en el gráfico, mayor es la presencia de esa taxonomía en las muestras.

En esta sección se puede presentar un solo gráfico de las taxonomías compartidas entre todas las muestras, y en el caso de que el usuario haya ingresado grupos, se desplegaran además los gráficos por cada grupo. 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.

images/app/results/basicStatistics.png

Figura 3.23. Gráfico de similitud (resultados)

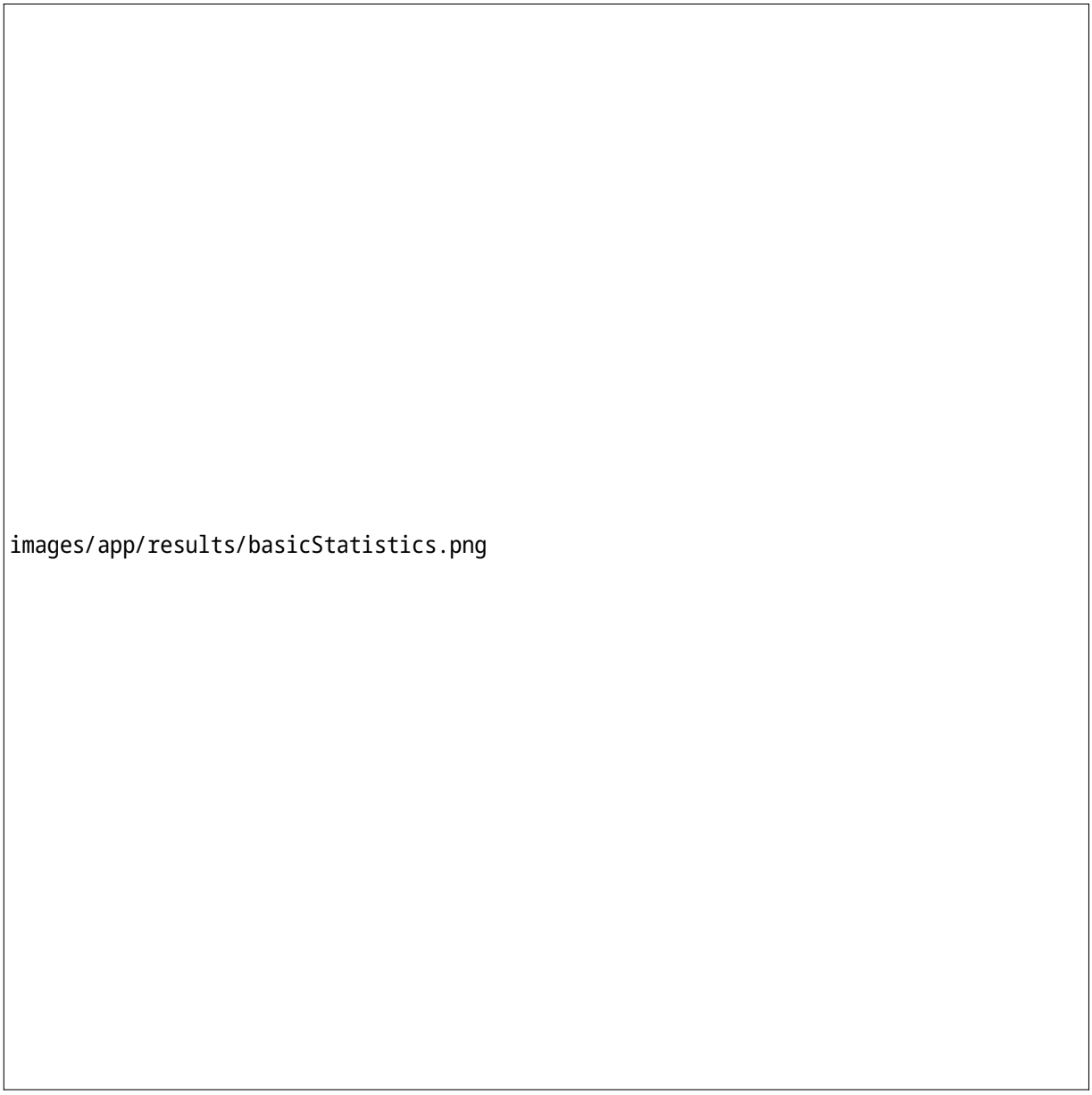


Figura 3.24. Gráfico de similitud: Tooltip asociado a las muestras (resultados)

En caso de que un grupo no tenga taxonomías compartidas entre si, se desplegará un mensaje indicando esto.

3.2.10 | Indices de diversidad

hacer

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.

3.2.11 | Predicción funcional

Al igual que en la sección de asignación taxonómica, en la parte superior se pueden visualizar tres pestañas *EC, KO y Pathways* que representan cada categoría funcional. Por defecto se presenta la información para la categoría de *Pathways*.

En la parte inferior izquierda se puede visualizar una tabla con la información de la predicción funcional obtenida mediante PICRUSt2 (EC, KO y Pathways) para cada muestra. En la parte superior de la tabla 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.

Al igual que en la sección de asignación taxonómica el usuario puede interactuar con las pestañas modificando el contenido de las tablas mediante la selección de las pestañas.

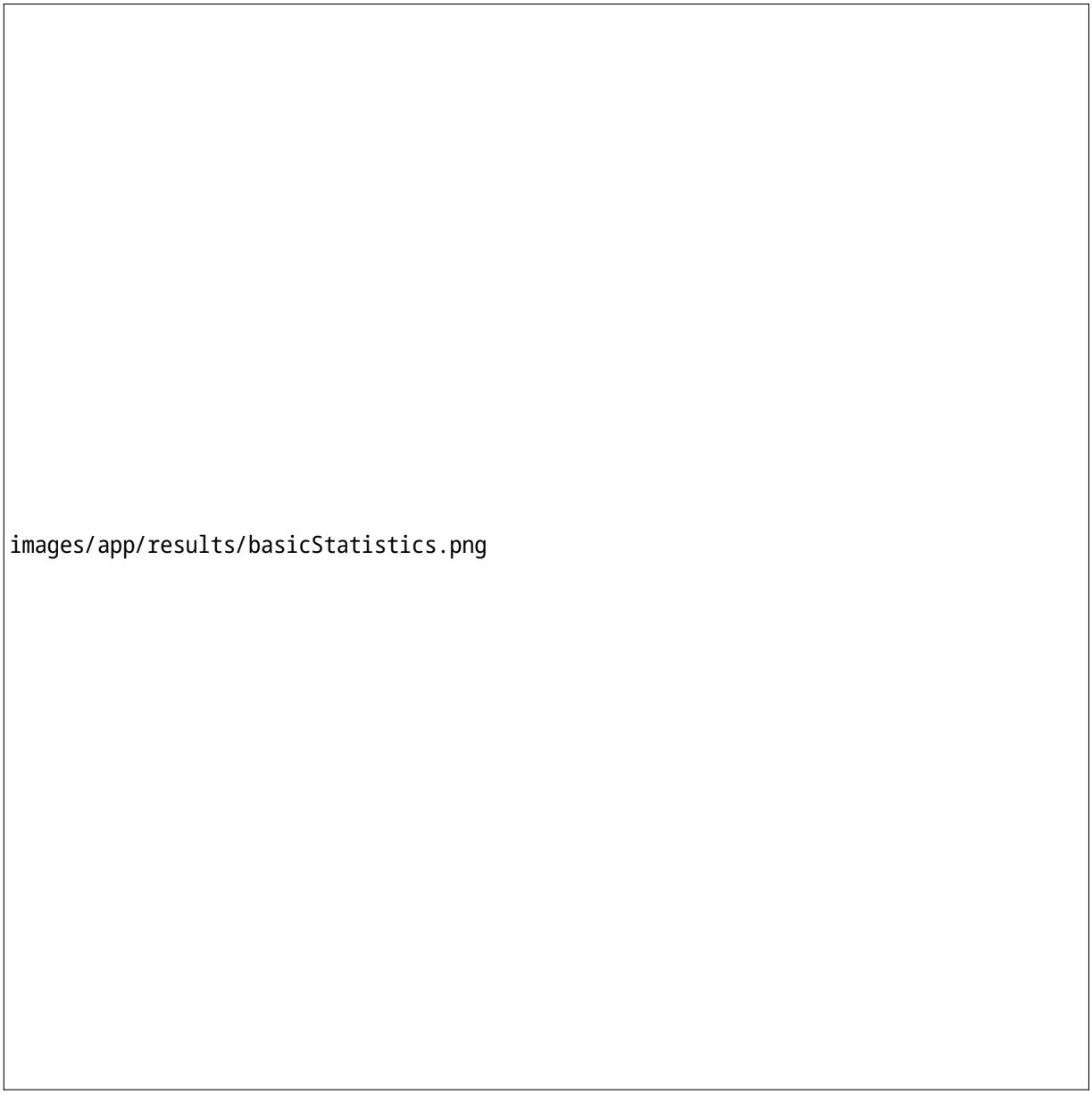


Figura 3.25. Gráfico de similitud: Tooltip asociado a las muestras (resultados)

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. A continuación se detallan los archivos:

- 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), en caso dehaber seleccionado predicción funcional dentro de los análisis.
- CSV con los valores del cálculo 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)

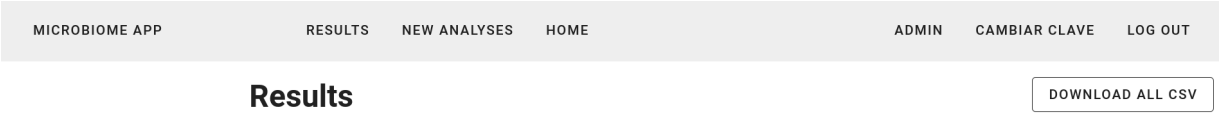


Figura 3.26. Gráfico de similitud: Tooltip asociado a las muestras (resultados)

3.2.12 | Integración del flujo de trabajo y aplicación web

Se diseño una base de datos no relacional utilizando PostgreSQL. Esta base de datos permite almacenar la información de los proyectos subidos por el usuario en la plataforma web y los resultados obtenidos en la ejecución del flujo de trabajo. La base de datos funciona como intermedio entre el flujo de trabajo y la base de datos, permitiendo que la aplicación web escriba la metadata asociada al proyecto para que el flujo de trabajo pueda ejecutarse, y permitiendo además que la aplicación web pueda leer los resultados escritos por el flujo de trabajo, para poder convertir esta información en tablas y gráficos que permitan al usuario visualizar los resultados de manera sencilla.

La base de datos esta compuesta por seis tablas:

- PlatformRoles: Almacena los roles de la plataforma (admin, basic user).
- Users: Almacena la información de los usuarios registrados en la plataforma.
- Projects: Almacena la información de los proyectos subidos por los usuarios.
- ProjectRoles: Almacena los roles de los usuarios dentro de los proyectos. Esto permite que un projecto y sus resultados puedan ser visualizados por varios usuarios a la vez.
- ProjectRolesUsers: Almacena la relación entre los usuarios y los roles de los proyectos. Permitiendo a los usuarios tener diferentes roles en diferentes proyectos (solo lectura, edición, eliminar).
- Results: Almacena los resultados obtenidos en la ejecución del flujo de trabajo.
- PipelineDefaultParams: Almacena los parámetros por defecto de cada versión del flujo de trabajo.

añadir cardinalidad

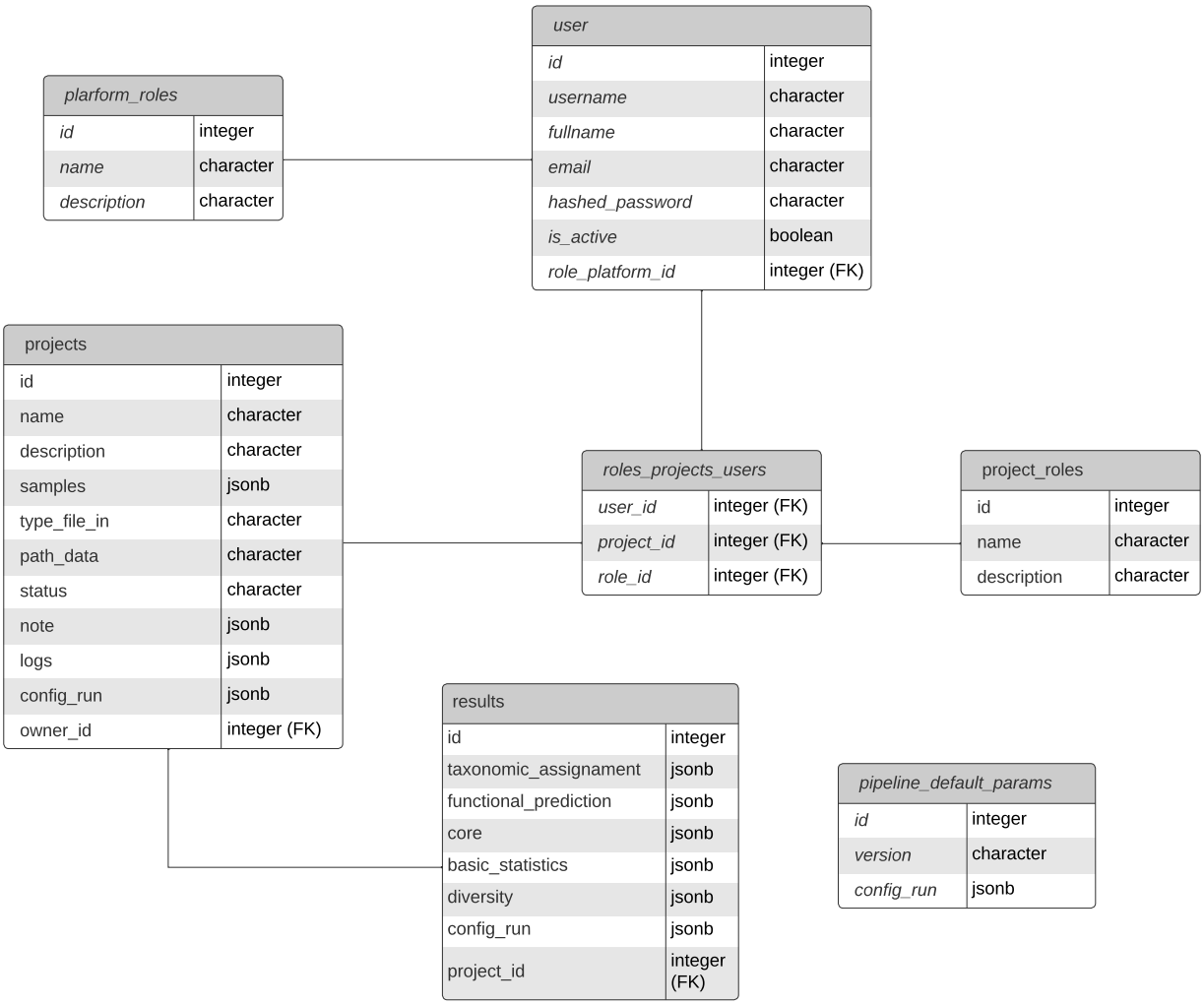


Figura 3.27. Base de datos

Las tablas de Resultados y Proyectos estan compuestas por campos JSONB que permiten almacenar datos en formato JSON mediante una representación binaria permitiendo que los datos sean indexados y consultados de manera eficiente.

Para la ejecución del flujo de trabajo se desarrollo un script en Python que escribe el archivo *params.yml* con la información del proyecto y los parámetros ingresados por el usuario en la plataforma (y guardados en la base de datos). Una vez que el pipeline finaliza su ejecución mediante un script en Python se almacenan los resultados en la base de datos.

backend

3.2.13 | 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

4.1 | Trabajos futuros

- clustering
- contan con cultivo de bacterias secuenciados que me permitan hacer validación del metodo
- Terminar de implementar las buenas normas de nf-core
- Gestión de cuentas
- Creación de cuentas
- Graficos cuando son mas de 30 muestras
- gestionar la unión de proyectos y metadata
- hacer documentación
- se ve mal la app con algunos zooms
- clustering con la nueva química

glosario

- PCR (Polymerase chain reaction o reacción en cadena de la polimerasa): Técnica de la biología molecular para hacer muchas copias a partir de un fragmento de ADN.
- 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. Gilbert, J. A., Blaser, M. J. *y col.* Current understanding of the human microbiome. *Nature medicine* **24,** 392-400 (2018).
2. Bahrndorff, S., Alemu, T., Alemneh, T., Lund Nielsen, J. *y col.* The microbiome of animals: implications for conservation biology. *International journal of genomics* **2016** (2016).
3. Berendsen, R. L., Pieterse, C. M. & Bakker, P. A. The rhizosphere microbiome and plant health. *Trends in plant science* **17,** 478-486 (2012).
4. 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).
5. Banerjee, S. & Van Der Heijden, M. G. Soil microbiomes and one health. *Nature Reviews Microbiology* **21,** 6-20 (2023).
6. Marco, M. L. Defining how microorganisms benefit human health. *Microbial Biotechnology* **14,** 35-40 (2021).
7. Fijan, S. Microorganisms with claimed probiotic properties: an overview of recent literature. *International journal of environmental research and public health* **11,** 4745-4767 (2014).
8. 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).
9. Hou, K., Wu, Z.-X. *y col.* Microbiota in health and diseases. *Signal transduction and targeted therapy* **7,** 1-28 (2022).
10. Ursell, L. K., Clemente, J. C. *y col.* The interpersonal and intrapersonal diversity of human-associated microbiota in key body sites. *Journal of Allergy and Clinical Immunology* **129,** 1204-1208 (2012).
11. Bitton, G. Role of microorganisms in biogeochemical cycles. *Wastewater Microbiology.,* 51-73 (1994).
12. 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).
13. 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).
14. Bickel, S. & Or, D. Soil bacterial diversity mediated by microscale aqueous-phase processes across biomes. *Nature Communications* **11,** 116 (2020).
15. Hu, J., Wei, Z. *y col.* Probiotic Pseudomonas communities enhance plant growth and nutrient assimilation via diversity-mediated ecosystem functioning. *Soil Biology and Biochemistry* **113,** 122-129 (2017).
16. Lemanceau, P., Blouin, M., Muller, D. & Moënne-Loccoz, Y. Let the core microbiota be functional. *Trends in Plant Science* **22,** 583-595 (2017).
17. Hardoim, P. R., Van Overbeek, L. S. *y col.* The hidden world within plants: ecological and evolutionary considerations for defining functioning of microbial endophytes. *Microbiology and molecular biology reviews* **79,** 293-320 (2015).
18. Vorholt, J. A. Microbial life in the phyllosphere. *Nature reviews microbiology* **10,** 828-840 (2012).
19. 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).
20. Tun, H. M., Konya, T. *y col.* Exposure to household furry pets influences the gut microbiota of infants at 3–4 months following various birth scenarios. *Microbiome* **5,** 1-14 (2017).
21. Azad, M. B., Konya, T. *y col.* 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).
22. Kates, A. E., Jarrett, O. *y col.* Household pet ownership and the microbial diversity of the human gut microbiota. *Frontiers in cellular and infection microbiology* **10,** 73 (2020).
23. Roslund, M. I., Puhakka, R. *y col.* Biodiversity intervention enhances immune regulation and health-

associated commensal microbiota among daycare children. *Science advances* **6,** eaba2578 (2020).
24. 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).
25. Olsen, G. J. & Woese, C. R. Ribosomal RNA: a key to phylogeny. *The FASEB journal* **7,** 113-123 (1993).
26. Reller, L. B., Weinstein, M. P. & Petti, C. A. Detection and identification of microorganisms by gene amplification and sequencing. *Clinical infectious diseases* **44,** 1108-1114 (2007).
27. 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).
28. López-Aladid, R., Fernández-Barat, L. *y col.* Determining the most accurate 16S rRNA hypervariable region for taxonomic identification from respiratory samples. *Scientific reports* **13,** 3974 (2023).
29. Patel, J. B. 16S rRNA gene sequencing for bacterial pathogen identification in the clinical laboratory. *Molecular diagnosis* **6,** 313-321 (2001).
30. Klindworth, A., Pruesse, E. *y col.* 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).
31. 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).
32. 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).
33. 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).
34. Didelot, X., Bowden, R., Wilson, D. J., Peto, T. E. & Crook, D. W. Transforming clinical microbiology with bacterial genome sequencing. *Nature Reviews Genetics* **13,** 601-612 (2012).
35. 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).
36. 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).
37. Watson, J. D. & Crick, F. H. Molecular structure of nucleic acids: a structure for deoxyribose nucleic acid. *Nature* **171,** 737-738 (1953).
38. Kumar, K. R., Cowley, M. J. & Davis, R. L. *Next-generation sequencing and emerging technologies* en *Seminars in thrombosis and hemostasis* (2024).
39. Bierman, G., Abadi, M. & Torgersen, M. *Understanding typescript* en *European Conference on Object-Oriented Programming* (2014), 257-281.
40. 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).
41. Crossley, B. M., Bai, J. *y col.* Guidelines for Sanger sequencing and molecular assay monitoring. *Journal of Veterinary Diagnostic Investigation* **32,** 767-775 (2020).
42. Lamoureux, C., Surgers, L. *y col.* Prospective comparison between shotgun metagenomics and sanger sequencing of the 16S rRNA gene for the etiological diagnosis of infections. *Frontiers in Microbiology* **13,** 761873 (2022).
43. Mardis, E. R. Next-generation DNA sequencing methods. *Annu. Rev. Genomics Hum. Genet.* **9,** 387-402 (2008).
44. Heather, J. M. & Chain, B. The sequence of sequencers: The history of sequencing DNA. *Genomics* **107,** 1-8 (2016).
45. Salipante, S. J., Kawashima, T. *y col.* 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).
46. 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).
47. 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).
48. He, Y., Zhou, B.-J. *y col.* Comparison of microbial diversity determined with the same variable tag sequence extracted from two different PCR amplicons. *BMC microbiology* **13,** 1-8 (2013).
49. Claesson, M. J., Wang, Q. *y col.* 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).
50. Szoboszlay, M., Schramm, L. *y col.* 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).
51. Benítez-Páez, A., Portune, K. J. & Sanz, Y. Species-level resolution of 16S rRNA gene amplicons sequenced through the MinION™ portable nanopore sequencer.

GigaScience **5,** s13742-016-0111-z. doi:10.1186/s13742-016-0111-z (ene. de 2016).

52. Amarasinghe, S. L., Su, S. *y col.* Opportunities and challenges in long-read sequencing data analysis. *Genome biology* **21,** 30 (2020).
53. Deamer, D., Akeson, M. & Branton, D. Three decades of nanopore sequencing. *Nature biotechnology* **34,** 518-524 (2016).
54. Cuber, P., Chooneea, D. *y col.* Comparing the accuracy and efficiency of third generation sequencing technologies, Oxford Nanopore Technologies, and Pacific Biosciences, for DNA barcode sequencing applications. *Ecological Genetics and Genomics* **28,** 100181 (2023).
55. 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).
56. 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. de 2020).
57. 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).
58. Curry, K. D., Wang, Q. *y col.* Emu: species-level microbial community profiling of full-length 16S rRNA Oxford Nanopore sequencing data. *Nature methods* **19,** 845-853 (2022).
59. Andrews, S. *y col. FastQC: a quality control tool for high throughput sequence data* 2010.
60. 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 (mayo de 2023).
61. Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34,** i884-i890 (2018).
62. 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).
63. Douglas, G. M., Maffei, V. J. *y col.* PICRUSt2 for prediction of metagenome functions. *Nature biotechnology* **38,** 685-688 (2020).
64. Segata, N., Izard, J. *y col.* Metagenomic biomarker discovery and explanation. *Genome biology* **12,** 1-18 (2011).
65. Dixon, P. VEGAN, a package of R functions for community ecology. *Journal of vegetation science* **14,** 927-930 (2003).
66. 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).
67. Di Tommaso, P., Chatzou, M. *y col.* Nextflow enables reproducible computational workflows. *Nature biotechnology* **35,** 316-319 (2017).
68. *Anaconda Software Distribution* ver. Vers. 23.9.0. 2024.
69. Kurtzer, G. M., Sochat, V. & Bauer, M. W. Singularity: Scientific containers for mobility of compute. *PloS one* **12,** e0177459 (2017).