Universidad Europea

Máster en Bioinformática

Evaluación de la Adherencia a los Principios FAIR en Bases de Datos de Proyectos de Secuenciación Metagenómica para Enfermedades Inflamatorias Intestinales

Autor: Osman Javier López Gualdrón

Tutor: María del Rocío González Soltero

Curso 2023-24

AGRADECIMIENTOS

Quiero comenzar agradeciendo a mi esposa, **Erika Alarcón**, por su constante apoyo y compañía. Erika, en cada una de esas noches en las que me faltaban las ideas y sentía que ya no podía avanzar, estuviste ahí para animarme y recordarme que el esfuerzo valía la pena. Gracias por estar siempre a mi lado.

A mi tutora, **María del Rocío González**, quiero agradecerle profundamente por su guía y dedicación a lo largo de este camino. Gracias por acompañarme en cada paso del proceso y por ayudarme aclarar las ideas en los momentos más desafiantes. Tu apoyo fue esencial para que este proyecto se pudiera concluir.

Resumen

Introducción: Las Enfermedades Inflamatorias Intestinales (EII), como la Enfermedad de Crohn (EC) y la Colitis Ulcerosa (CU), están asociadas con alteraciones en la microbiota intestinal. Este trabajo plantea la hipótesis de que las bases de datos de secuenciación metagenómica disponibles para el estudio de EII presentan limitaciones en cuanto a su adherencia a los principios FAIR, lo que afecta su aplicación en modelos de aprendizaje automático.

Objetivos: Evaluar las principales bases de datos de secuenciación metagenómica para EII con el fin de determinar su adherencia a los principios FAIR y su utilidad en modelos de aprendizaje automático.

Metodología: Se realizó una revisión sistemática utilizando el protocolo PRISMA para identificar bases de datos abiertas relevantes. Luego, se llevó a cabo un análisis bioinformático de cuatro bases de datos identificadas (GMrepo, MGnify, HumanMetagenomeDB, HPMCD), utilizando Python y herramientas estadísticas para evaluar la calidad de los datos, y la adherencia a los principios FAIR fue evaluada mediante un cuestionario implementado en RedCap por la Universidad Europea de Madrid.

Resultados: Las bases de datos mostraron buena accesibilidad e interoperabilidad, pero se identificaron deficiencias en la estandarización de metadatos y reutilización, lo que limita su integración en modelos de aprendizaje automático. MGnify fue la más destacada en diversidad de muestras, pero todas presentaron una limitada diversidad étnica.

Conclusiones: Aunque las bases de datos de secuenciación metagenómica para EII han avanzado en accesibilidad e interoperabilidad, las deficiencias en la reutilización y estandarización limitan su aplicación en estudios más amplios y en modelos predictivos.

Palabras clave: Enfermedad de Crohn, Colitis Ulcerosa, secuenciación metagenómica, principios FAIR, aprendizaje automático

Abstract

Introduction: The inflammatory bowel diseases (IBD) such as Chohn's Disease (CD) and Ulcerative Colitis (UC), these ones are associated with alterations in the gut microbiota. This work proposes the hypothesis that the available metagenomic sequencing databases for the IBD studies have limitations regarding its adherence to FAIR principles, which affects their application in machine learning models.

Objectives: Evaluate the main metagenomic sequencing databases for IBD in order to determine their adherence to FAIR principles and their usefulness in machine learning models.

Methodology: A systematic review was conducted using the PRISMA protocol to identify relevant open databases. Then, a bioinformatics analysis of four identified databases (GMrepo, MGnify, HumanMetagenomeDB, HPMCD) was carried out using Python and statistical tools to assess data quality and the adherence to FAIR principles was evaluated through a questionnaire implemented in RedCap by the European University of Madrid

Results: The databases showed good accessibility and interoperability, but deficiencies were identified in the metadata standardization and reusability, this limits their integration into machine learning models. MGnifiy was the most outstanding in sample diversity, but all databases showed limited ethnic diversity.

Conclusions: Although metagenomic sequencing databases for IBD have advance in accessibility and interoperability, the deficiencies in reuse and standardization limit their application in broader studies and predictive models.

Keywords: Crohn's Disease, Ulcerative Colitis, metagenomic sequencing, FAIR principles, machine learning.

INDICE DE CONTENIDOS

1 Introducción 9

1.1 Desafíos en la utilización de datos de secuenciación metagenómica 11

1.1.1 Calidad de los datos metagenómicos 11

1.2 Complejidad y heterogeneidad de las EII 12

1.3 Justificación 13

2 Pregunta de investigación, hipótesis y objetivos 14

2.1 Pregunta de investigación 14

2.2 Hipótesis 15

2.3 Objetivo general y objetivos específicos 15

3 Metodología 15

3.1 Revisión bibliográfica sistemática 15

3.2 Análisis bioinformático 17

3.2.1 Identificación y Extracción de Datos 17
3.2.2 Análisis Descriptivo y Estadístico 17

3.3 Evaluación de la adherencia a los principios FAIR 18

4 Resultados 18

4.1 Identificación y clasificación de las bases de datos 19

4.1.1 Característica de artículos seleccionados de la revisión sistemática 21
4.1.2 Protocolo de extracción de datos 23

4.2 Análisis descriptivo de las bases de datos identificadas 24

4.2.1 Distribución del número de muestras 25
4.2.2 Frecuencia de actualización 25
4.2.3 Diversidad de las poblaciones 25
4.2.4 HPMCD (Human Pan-Microbial Communities Database) 27
4.2.5 GMrepoV2 28
4.2.6 MGnify 31
4.2.7 HMgDB 33

4.3 Análisis de adherencia a los principios FAIR en bases de datos metagenómicas 34

5 Discusión 37

6 Conclusiones 39

7 Bibliografía 41

8 Anexo 1 47

9 Anexo 2: Listado Cotejo IA 54

INDICE DE FIGURAS

Figura 1.Figura 1. Respuesta inmune. Creado en BioRender (Abraham & Cho, 2009). 10

Figura 2. PRISMA selección de artículos de la revisión bibliográfica. Elaboración propia creada con covidence.com 20

Figura 3. Análisis descriptivo de las bases de datos identificadas. Creado en Python librería Matplotlib 26

Figura 4. HPMCD (Human Pan-Microbial Communities Database) 28

Figura 5. Distribución del nivel taxonómico. Creado en Python librería Matplotlib 30

Figura 6. Distribución del nombre del bioma MGnify. Creado con Python libería Matplotlib ... 32

Figura 7. Distribución de diagnóstico HMgDB. Creado con Python libería Matplotlib 34

Figura 8. Análisis de adherencia a los principios FAIR en bases de datos metagenómicas. Creado con Python libería Matplotlib 36

ÍNDICE DE TABLAS

Tabla 1. Objetivo general y específicos. Elaboración propia 15

Tabla 2. Criterios de inclusión y exclusión para la revisión bibliográfica. Elaboración propia 16

Tabla 3. Característica de artículos seleccionados de la revisión sistemática. Elaboración propia..... 22

Tabla 4. Protocolo de extracción de datos. Elaboración propia 24

Acrónimos

EII: Enfermedad Inflamatoria Intestinal

EC: Enfermedad de Crohn

CU: Colitis Ulcerosa

FAIR: Findable, Accessible, Interoperable, Reusable

HMgDB: Human Metagenome Database

HPMCD: Human Pan-Microbial Communities Database

GMrepo: Gut Microbiota Repository

MGnify: Metagenomics Analysis Platform

Q30: Calidad superior a 30 en secuenciación genética

ML: Machine learning

IL-12: Interleucina 12

IL-23: Interleucina 23

1 Introducción

Las Enfermedades Inflamatorias Intestinales (EII) comprenden un grupo de trastornos crónicos que afectan el tracto gastrointestinal, destacando principalmente la EC (EC) y la CU (CU). Aunque ambas condiciones comparten características clínicas comunes, se diferencian significativamente en su presentación anatómica y patogénesis. La EC se caracteriza por una inflamación transmural que puede afectar cualquier parte del tracto gastrointestinal, desde la boca hasta el ano, y presenta lesiones discontinuas conocidas como "skip lesions". Por otro lado, la CU se limita al colon y recto, con una inflamación continua que afecta principalmente la mucosa y submucosa (Hemmer, Forest, Rath, & Bowman, 2023).

Las EII han mostrado un aumento en su prevalencia tanto en los Estados Unidos como en el Reino Unido. En los Estados Unidos, aproximadamente 1.6 millones de personas viven con EII, con 70,000 nuevos casos diagnosticados anualmente (Crohn's & Colitis Foundation of America, 2014). En el Reino Unido, la prevalencia de EII aumentó en un 33.8% entre 2006 y 2016, lo que destaca la creciente carga de estas enfermedades (Freeman et al., 2021).

La patogénesis de la EII es multifactorial, involucrando una interacción entre factores genéticos, ambientales y microbianos (Crohn's & Colitis Foundation of America, 2014), teniendo como resultado una respuesta inmune alterada. La figura 1 ilustra cómo esta alteración conduce a un aumento de la permeabilidad epitelial y una inflamación crónica mediada por citocinas proinflamatorias como el Factor de Necrosis Tumoral Alfa, L-12 e IL-23, las cuales son importantes en el desarrollo del ciclo inflamatorio en el tejido intestinal (Abraham & Cho, 2009). Asociado a la alteración de la permeabilidad epitelial y la inflamación crónica, la pérdida de la capa de moco y la disfunción de la barrera epitelial permiten la entrada de patógenos en la lámina propia, exacerbando la inflamación y llevando a una expansión significativa de este tejido (Abraham & Cho, 2009). La identificación de más de 160 genes asociado con EII ha permitido una mejor comprensión de estos mecanismos subyacentes, destacando la importancia de la predisposición genética y el papel del microbioma intestinal en la inflamación crónica (Crohn's & Colitis Foundation of America, 2014). Por otra parte, los factores ambientales

como la dieta occidental, caracterizada por un alto consumo de grasa y alimentos procesados, han sido vinculados con un mayor riesgo de desarrollo de EII (Manski et al., 2024).

Estudios recientes han demostrado que no solo la composición del microbioma, sino también las características del metaboloma juegan un papel clave en la progresión de estas enfermedades. Análisis integrados de múltiples cohortes han revelado que alteraciones específicas en el microbioma y el metaboloma están estrechamente asociadas con la severidad de las EII, sugiriendo que estos biomarcadores pueden servir como herramientas diagnósticas y para el monitoreo de la enfermedad (Ning et al., 2023).

La diversidad y composición del microbioma intestinal han surgido como factores clave en la diferenciación y diagnóstico de las EII. Estudios recientes han demostrado que la diversidad microbiana se reduce significativamente en pacientes con CU, con esta reducción siendo más pronunciada en pacientes con EC (Kang et al., 2023). Estos cambios en el microbioma no solo reflejan la disbiosis, sino que también están asociados con alteraciones funcionales que contribuyen a la patogénesis de la enfermedad, lo que resalta el potencial del microbioma como un marcador diagnóstico y una diana terapéutica en las EII (Shan et al., 2022).

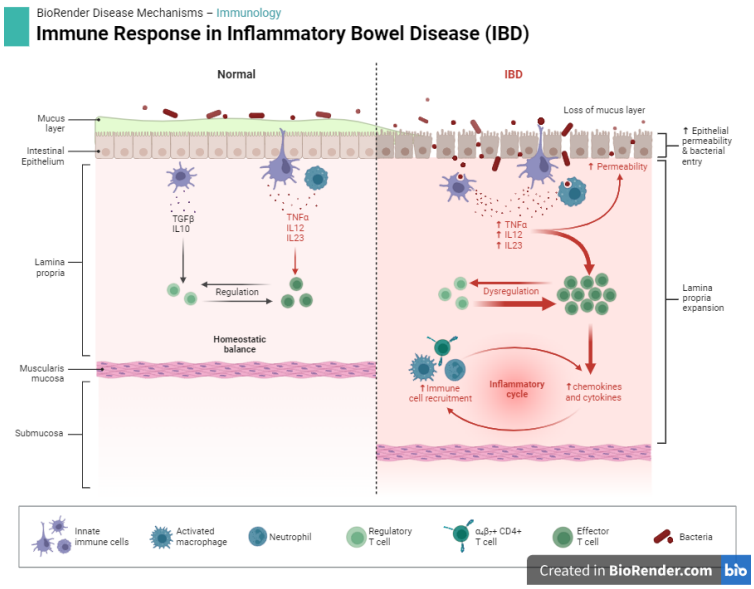


Figura 1.Figura 1. Respuesta inmune. Creado en BioRender (Abraham & Cho, 2009).

El impacto de factores externos, como la dieta occidental, también ha sido ampliamente estudiado en relación con las EII. Este tipo de dieta, caracterizada por un alto consumo de grasas y alimentos procesados, se asocia con un aumento en la incidencia de EII, particularmente en regiones donde anteriormente estas condiciones eran raras (Manski et al., 2024). En contraste, enfoques dietéticos específicos, como la nutrición enteral y la dieta mediterránea, han mostrado beneficios en la gestión de los síntomas y en la inducción de remisión en pacientes con EII (Manski et al., 2024).

El diagnóstico y manejo de las EII también han evolucionado con el uso de técnicas avanzadas como la endoscopia y la imagenología de sección transversal, que son esenciales para la evaluación de la extensión de la enfermedad y la identificación de complicaciones. Además, las terapias biológicas y los inmunomoduladores han revolucionado el tratamiento de las EII, ofreciendo nuevas esperanzas para la inducción y mantenimiento de la remisión (Veauthier & Hornecker, 2018).

1.1 Desafíos en la utilización de datos de secuenciación metagenómica

El uso de secuenciación metagenómica en el estudio de las EII ha permitido avances significativos en la comprensión de la composición y función del microbioma intestinal (Valles-Colomer et al., 2016). Sin embargo, el manejo y análisis de estos datos presentan desafíos importantes, particularmente en términos de calidad, accesibilidad y adherencia a los principios FAIR (Encontrable, Asequible, Interoperabilidad, reutilizable), estos garantizan que los datos no solo sean útiles en el corto plazo, sino que también conserven su valor y aplicabilidad en el futuro, acelerando los avances en el campo de la secuenciación metagenómica y las EII (Valles-Colomer et al., 2016).

1.1.1 Calidad de los datos metagenómicos

El control de calidad es esencial para evitar que secuencias de baja calidad, contaminantes o artefactos comprometan el análisis. La mala calidad de los datos puede llevar a errores en el ensamblaje y a conclusiones incorrectas, afectando la reproducibilidad y fiabilidad de los estudios (Schmieder & Edwards, 2011). La necesidad de filtrar las secuencias según su longitud, contenido de Guanina – Citosina y puntuaciones de calidad es crucial para mejorar tareas como la alineación de secuencias,

identificación de variantes genéticas, anotación funcional y la interpretación de los datos, reduciendo el ruido en los datos y asegurando que los datos sean utilizables y fiables (Schmieder & Edwards, 2011).

La accesibilidad de los datos y adherencia a los principios FAIR representan un desafío considerable. A pesar de que existen plataformas web como Galaxy, MG-RAST, y EBI Metagenomics que democratizan el acceso al análisis de datos metagenómicos, la falta de estandarización entre estas herramientas limita su interoperabilidad, lo que complica la reutilización de los datos en diferentes estudios (Achudhan et al., 2024). La automatización en estas plataformas puede facilitar el análisis, pero también restringe la flexibilidad necesaria para adaptar los análisis a necesidades específicas, lo que es un obstáculo para su uso amplio y efectivo.

Aunque la metagenómica ofrece una gran cantidad de datos, la falta de estandarización en la recolección de muestras y en los métodos de procesamiento de datos complica su integración y comparación entre diferentes estudios, lo que afecta la interoperabilidad y la reutilización de los datos (Valles-Colomer et al., 2016). Es fundamental desarrollar enfoques estandarizados que permitan una interpretación efectiva de los datos para descubrir patrones clínicamente relevantes y garantizar que los datos metagenómicos cumplan con los principios FAIR.

1.2 Complejidad y heterogeneidad de las EII

En el contexto de las EII, la complejidad inherente de estas patologías, combinada con su causalidad multifactorial, plantea serios desafíos para la creación de modelos predictivos robustos basados únicamente en datos metagenómicos. La integración de estos datos con otras fuentes ómicas, como la transcriptómica y la proteómica, es esencial para capturar toda la gama de factores biológicos subyacentes a la progresión y severidad de las EII. Este enfoque no solo permite una mejor identificación de subtipos moleculares, como se ha demostrado en estudios recientes, sino que también contribuye a la personalización de las terapias, adaptando el tratamiento a los perfiles moleculares específicos de los pacientes (Gonzalez et al., 2022; Preto et al., 2024).

Sin embargo, el uso efectivo de estos datos para la medicina de precisión requiere un esfuerzo considerable en términos de accesibilidad y reutilización de la información. Aquí es donde los principios FAIR juegan un papel crucial. Al asegurar que los datos multi-ómicos sean fáciles de encontrar, accesibles, interoperables y reutilizables, no solo se optimiza la investigación, sino que también se garantiza que los modelos predictivos generados sean precisos y clínicamente útiles en diversos entornos (Seyed Tabib et al., 2020; Imhann et al., 2019). La adherencia a estos principios es un requisito esencial para mejorar la precisión diagnóstica y pronóstica de las EII, como ha sido subrayado en proyectos de gran escala como el 1000IBD, que han hecho una significativa contribución a la medicina personalizada a través de la integración de datos multi-ómicos (Imhann et al., 2019).

1.3 Justificación

El presente trabajo se centra en la evaluación crítica de las bases de datos de secuenciación metagenómica utilizadas para el estudio de la EC, CU y otras EII. Esta evaluación es esencial debido al creciente uso de modelos de Machine Learning (ML) en la investigación biomédica, donde la calidad de los conjuntos de datos es clave para el éxito de dichos modelos. En estudios recientes, se ha demostrado que la construcción y etiquetado de conjuntos de datos con alta precisión, que incluyan variables clínicas y demográficas, son cruciales para el desarrollo de modelos robustos y aplicables en la práctica clínica (Syed et al., 2024; Javaid et al., 2021). Sin embargo, la falta de integración funcional de datos multi-ómicos sigue siendo una barrera importante para lograr avances clínicos significativos en el manejo de las EII (Fiocchi, 2023).

Uno de los principales desafíos en la utilización de estos datos es evitar la fuga de información entre los conjuntos de entrenamiento y prueba, lo que podría introducir sesgos y comprometer la validez de los modelos predictivos. La construcción de modelos de ML ha permitido identificar biomarcadores clave y optimizar el diagnóstico temprano y el tratamiento personalizado de la EII. Por ejemplo, en estudios recientes, se ha demostrado la efectividad de técnicas como el Random Forest para predecir la respuesta a terapias biológicas en pacientes con EII, mejorando la precisión diagnóstica y

reduciendo los costos asociados (Syed et al., 2024; Ahmad et al., 2023; Javaid et al., 2021).

En el contexto de las EII, la representatividad de los datos es otro desafío significativo. Estudios han mostrado que la gran variabilidad biológica y clínica entre los pacientes complica la creación de modelos predictivos generalizables, lo que subraya la necesidad de integrar múltiples capas de datos ómicos para abordar esta heterogeneidad (Fiocchi, 2023). Además, la actualización continua de los datos es fundamental para asegurar que los modelos reflejen con precisión los fenotipos clínicos actuales y no se vuelvan obsoletos. Proyectos como el 1000IBD han mostrado cómo la aplicación de estos principios ha facilitado la identificación de subtipos moleculares y la optimización de las terapias personalizadas, demostrando que la interoperabilidad y la reutilización efectiva de los datos son esenciales para la integración en modelos de ML y la reproducibilidad científica (Imhann et al., 2019; Wilkinson et al., 2016; Krassowski et al., 2020).

2 Pregunta de investigación, hipótesis y objetivos

2.1 Pregunta de investigación

¿Cuáles son las principales bases de datos de secuenciación metagenómica para el estudio de EC, CU y en general las Enfermedades Inflamatorias Intestinales? Además, ¿Estas bases de datos cumplen con los principios FAIR para que puedan ser utilizadas en modelos de aprendizaje automático?

2.2 Hipótesis

Las principales bases de datos de secuenciación metagenómica para el estudio de las EII, no cumplen en su totalidad los principios FAIR, porque presentan dificultades en cuanto a la interoperabilidad y reutilización, limitando su aplicación en modelos de ML.

2.3 Objetivo general y objetivos específicos

Objetivo	Descripción
Objetivo General	Evaluar los conjuntos de datos de secuenciación metagenómica de EC, CU y en general EII disponibles en bases de datos abiertas para determinar su utilidad en modelos de aprendizaje automático, enfocándose en la integridad y adherencia a los principios FAIR.
Objetivo Específico 1	Identificar y clasificar las bases de datos abiertas que contienen datos de secuenciación metagenómica de EC, CU y en general las enfermedades inflamatorias intestinales, describiendo las características y el volumen de los datos disponibles.
Objetivo Específico 2	Realizar un análisis estadístico descriptivo de los datos de secuenciación metagenómica en las bases de datos seleccionadas, enfocándose en variables clave como estado de la enfermedad, género, etnicidad, nivel taxonómico, bioma, centro de investigación y métricas de calidad de secuenciación.
Objetivo Específico 3	Analizar la adherencia de los conjuntos de datos a los principios FAIR, a través de la herramienta RedCap desarrollada por la Universidad Europea de Madrid.

Tabla 1. Objetivo general y específicos. Elaboración propia

3 Metodología

Este trabajo se llevó a cabo mediante un enfoque mixto que combinó una revisión bibliográfica con un análisis bioinformático detallado. El propósito de esta metodología fue identificar y analizar las bases de datos de secuenciación metagenómica relevantes para las EII, como la EC y la CU, con el fin de evaluar la integridad y adherencia a los principios FAIR.

3.1 Revisión bibliográfica sistemática

Se realizó utilizando el protocolo PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) para asegurar la transparencia en el proceso de

selección de estudios. El análisis y la gestión de los artículos se llevaron a cabo mediante la plataforma Covidence, la cual facilita la eliminación automática de duplicados, permite un trabajo colaborativo y una revisión por pares de forma eficiente, también simplifica la organización de los estudios y su clasificación de acuerdo a los criterios de inclusión y exclusión predefinidos (Cleo et al., 2019).

Para la revisión de artículos relevantes se consultó la base de datos PubMed, utilizando las palabras clave: *Metagenomics*, *Metadata*, *Ontologies*, *Sequence read archive* y *Genomic databases*. La estrategia de búsqueda fue la siguiente: *(Metagenomics[MeSH] OR "metagenomics") AND ("metadata" OR "ontologies" OR "sequence read archive" OR "genomic databases")*.

Se establecieron criterios de inclusión y exclusión para asegurar la relevancia y calidad de los estudios seleccionados que se muestran en la tabla 2:

Criterios de Inclusión	Criterios de Exclusión
Investigaciones originales, revisiones sistemáticas, o meta-análisis, estudios de cohorte, casos y controles, transversales, series de caso que involucren metagenómica.	Editoriales, cartas al editor, comentarios y resúmenes de conferencias que no proporcionen datos completos.
Estudios que describen el desarrollo o la utilización de bases de datos abiertas relacionadas con la secuenciación metagenómica.	Publicaciones duplicadas.
Artículos que mencionen explícitamente bases de datos de metagenómica.	Estudios que no mencionan metagenómica o enfermedades intestinales inflamatorias*.
Estudios centrados en enfermedades inflamatorias intestinales, específicamente la enfermedad de Crohn y colitis ulcerosa.	Estudios que no discutan bases de datos.
Estudios que discutan o utilicen bases de datos de acceso libre y público. Artículos escritos en inglés	Artículos que no hagan referencia al acceso a bases de datos o que mencionen bases de datos de acceso restringido.

Tabla 2. Criterios de inclusión y exclusión para la revisión bibliográfica. Elaboración propia

La estrategia de búsqueda no incluyó términos específicos para la EC y CU para asegurar una cobertura más amplia y evitar la exclusión de estudios relevantes que, aunque no mencionen explícitamente estas enfermedades en sus títulos o resúmenes, podrían contener datos útiles para el análisis de enfermedades inflamatorias intestinales. Al ampliar el alcance de la búsqueda inicial, se pudo identificar un mayor

número de estudios potencialmente relevantes, que luego fueron evaluados más detalladamente según los criterios específicos de inclusión y exclusión.

3.2 Análisis bioinformático

3.2.1 Identificación y Extracción de Datos

Se identificaron cuatro bases de datos relevantes para el estudio de la secuenciación metagenómica en EII: GMrepo, MGnify, HumanMetagenomeDB, y HPMCD. Para cada base de datos, se desarrolló un protocolo de extracción de datos que incluyó las siguientes variables: Nombre de la base de datos, descripción, enlace, número de muestras, técnicas de secuenciación, enfermedad estudiada, frecuencia de actualización y diversidad de las poblaciones representadas.

3.2.2 Análisis Descriptivo y Estadístico

Con el protocolo de extracción de datos desarrollado, se realizó un análisis descriptivo y estadístico de las bases de datos utilizando el lenguaje de programación Python, junto con las librerías pandas y matplotlib. La librería panda se empleó para la manipulación y análisis de los datos, mientras que matplotlib se utilizó para la generación de las gráficas.

El análisis se enfocó en la distribución del número de muestras, la frecuencia de actualización de las bases de datos y la diversidad de las poblaciones representadas. Además, se realizó un análisis detallado de variables clave como el estado de la enfermedad, género, etnicidad, nivel taxonómico y centro de investigación.

Para la base de datos HMgDB se evaluó la calidad de la secuenciación utilizando gráficos de boxplot, los cuales permitieron visualizar la distribución de diversas métricas, como la cantidad de secuencias, el número de pares de bases, la longitud promedio de las secuencias y la calidad superior a 30 (Q30), esta última es una métrica de calidad de secuenciación que indica que la probabilidad de que una base esté mal identificada es de 1 en 1000 (Singular Genomics, 2023).

3.3 Evaluación de la adherencia a los principios FAIR

Para cumplir con este objetivo, se realizó un análisis exhaustivo de la adherencia de los conjuntos de datos metagenómicos utilizando un formulario que incluía una serie de preguntas dirigidas a evaluar los distintos aspectos de los principios FAIR en cada base de datos. Este formulario fue desarrollado por la Universidad Europea de Madrid en RedCap (plataforma de captura de datos que permite la creación de encuestas y formularios personalizados) (Gonzalez Soltero et al., 2024). Este proceso involucró la evaluación de bases de datos identificadas en la revisión bibliográfica, tales como GMrepo v2, HPMCD, MGnify (HumGut) y HumanMetagenomeDB.

- Se implementó un cuestionario estructurado en REDCap para evaluar los principios FAIR.
- Cada base de datos respondió preguntas sobre la disponibilidad de identificadores, metadatos, y derechos de uso.
- Los resultados fueron recopilados automáticamente en formato CSV por la plataforma REDCap.
- El archivo CSV fue procesado y analizado usando Python para agrupar y visualizar los datos.
- Se utilizaron gráficos para identificar patrones de cumplimiento y áreas de mejora respecto a los principios FAIR.
- Para cada pregunta, se generaron gráficos de barras que mostraban la distribución de respuestas entre las bases de datos.
- Se calculó la frecuencia de cumplimiento de cada principio, permitiendo identificar tendencias generales en la adherencia a los principios FAIR.
- Los datos recopilados en formato CSV y el código utilizado para el análisis se encuentran disponibles en https://github.com/oslopez081/IBD-Data-FAIR-Principles/blob/main/README.md

4 Resultados

Los resultados se presentarán en dos secciones: La primera parte, corresponde a los resultados obtenidos tras la revisión sistemática y en la segunda parte abarca los hallazgos correspondientes al análisis bioinformático de cada una de las bases de datos identificadas.

4.1 Identificación y clasificación de las bases de datos

El proceso de selección de estudios se llevó a cabo utilizando la plataforma Covidence, siguiendo las etapas de identificación, cribado, elegibilidad e inclusión, conforme al diagrama de flujo PRISMA. En la fase de identificación, se localizaron 72 estudios en la base de datos PubMed, sin hallar referencias adicionales de otras fuentes, búsquedas de citas ni literatura gris. Durante el cribado inicial, se evaluaron los 72 estudios identificados. En esta primera etapa de evaluación por títulos y resúmenes, se excluyeron 31 estudios, dejando 41 estudios para la etapa siguiente. En esta segunda fase, cinco estudios fueron descartados por no discutir ni utilizar bases de datos y 31 por no mencionar metagenómica ni EII. Finalmente, se incluyeron 5 estudios en la revisión. La revisión sistemática llevada a cabo permitió identificar y catalogar varias bases de datos abiertas que contienen datos de secuenciación metagenómica relacionados con EII.

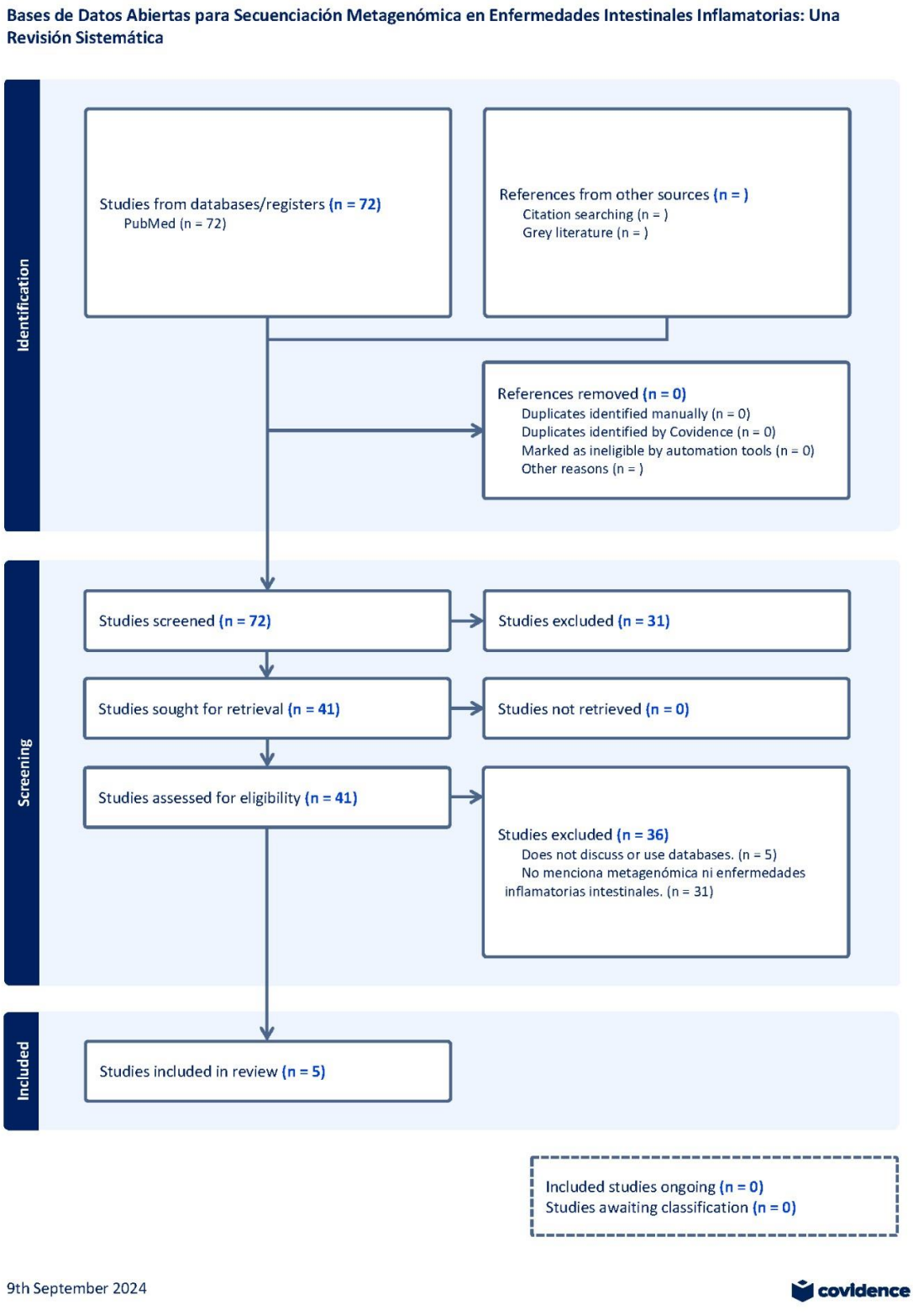


Figura 2. PRISMA selección de artículos de la revisión bibliográfica. Elaboración propia creada con covidence.com

4.1.1 Característica de artículos seleccionados de la revisión sistemática

Título del estudio, ID, Autor principal	Banco de datos	País	Fecha	Diseño de estudio	Financiamiento	Conflictos	Población	Intervención	Resultados
GMrepo v2: A Curated Human Gut Microbiome Database with Special Focus on Disease Markers and Cross-dataset Comparison. DOI: 10.1093/nar/gkab123	GMrepo v2	China	2020	Estudio de base de datos	Fundación Nacional de Ciencias Naturales de China	No hay conflictos de interés	Datos de 71,642 muestras del microbioma intestinal humano, abarcando 353 proyectos.	Recolección de datos metagenómicos del microbioma intestinal humano, anotación de los conjuntos de datos metagenómicos, identificación de marcadores microbianos asociados a enfermedades.	El conjunto de datos incluye marcadores de enfermedades como cáncer colorrectal, enfermedad de Crohn y obesidad. Incrementa la comparabilidad entre estudios mediante la curación de metadatos y mejora el acceso a los conjuntos de datos.
HPMCD: The Database of Human Microbial Communities from Metagenomic Datasets and Microbial Reference Genomes. DOI: 10.1093/nar/gkz744	HPMCD	Alemania	2020	Estudio de base de datos	Helmholtz Association (Alemania), DFG (Fundación Alemana de Investigación)	No hay conflictos de interés	Datos de más de 1,800 muestras metagenómicas humanas provenientes de muestras fecales.	Curación manual de datos metagenómicos intestinales humanos y organización basada en genomas de referencia microbiana.	Clasificación taxonómica a nivel de especies y cepas, y acceso a datos para análisis funcional de secuencias. Base de datos útil para estudiar la microbiota intestinal humana y comparar comunidades microbianas.

HumGut: A Comprehensive Human Gut Prokaryotic Genomes Collection Filtered by Metagenome Data. DOI: 10.1038/s41587-021-00923-0	HumGut	Noruega	2021	Estudio de base de datos	University of Oslo, Research Council of Norway	No hay conflictos de interés	Más de 30,000 genomas procariotas derivados de muestras metagenómicas del intestino humano.	Curación y agrupación de genomas procariotas del intestino humano utilizando muestras metagenómicas de todo el mundo.	La colección proporciona una alta resolución taxonómica para análisis de microbiomas intestinales, con el 97.5% de identidad entre secuencias, útil para la clasificación de un mayor porcentaje de lecturas en estudios del microbioma intestinal.
HumanMetagenomeDB: A Public Repository of Curated and Standardized Metadata for Human Metagenomes. DOI: 10.1093/nar/gkaa1031	HumanMetagenomeDB	Alemania	2020	Estudio de base de datos	National Key R&D Program of China, Helmholtz Association	No hay conflictos de interés	69,822 metagenomas humanos de 58 países, distribuidos en 580 estudios independientes.	Curación y estandarización de metadatos de metagenomas humanos, permitiendo la búsqueda según características del hospedador (edad, sexo, IMC, diagnóstico, etc.) y secuenciación.	La base de datos estandariza 203 atributos y facilita la reutilización de metadatos humanos para análisis comparativos. Permite identificar patrones microbiológicos en enfermedades como el cáncer y la enfermedad de Crohn.
GMrepo: A Database of Curated and Consistently Annotated Human Gut Metagenomes. DOI: 10.1093/nar/gkz764	GMrepo	China	2020	Estudio de base de datos	National Key Research and Development Program of China, National Natural Science Foundation of China	No hay conflictos de interés	58,903 muestras del microbioma intestinal humano de 253 proyectos, incluyendo 17,618 metagenomas y 41,285 amplicones.	Curación manual de metadatos y anotación consistente de los contenidos microbianos de muestras intestinales humanas, organizadas según fenotipos de salud y enfermedad.	Base de datos que permite realizar consultas biológicas complejas sobre la relación del microbioma intestinal con 92 fenotipos, incluidos cáncer colorrectal, enfermedades inflamatorias intestinales, entre otras. Facilita la reutilización de datos metagenómicos humanos.

Tabla 3. Característica de artículos seleccionados de la revisión sistemática. Elaboración propia

Cabe resaltar que de los artículos *GMrepo V2: A Curated Human Gut Microbione Database with Special Focus on Disease Markers and Cross-dataset Comparison* y *GMrepo: A Database of Curated and Consistently Annotated Human Gut Metagenomes,* hacen referencia a la misma base de datos, GMrepo, pero con enfoques diferentes. El primer artículo presenta una versión más reciente de la base de datos y la identificación de marcadores microbianos de enfermedades como el cáncer colorrectal y la EC. por otro lado, el segundo artículo describe la creación original de GMrepo, enfocándose en la curación manual de metadatos y la anotación consistente de secuencias microbianas provenientes de muestras del intestino humano.

Inicialmente, se tenía previsto realizar un análisis detallado de la base de datos HumGut, dada su relevancia en la colección de genomas procariotas del intestino humano sano. Sin embargo, al avanzar en la revisión de los estudios seleccionados, se observó que los datos metagenómicos presentes en HumGut se encuentran mayoritariamente disponibles a través del repositorio MGnify, que alberga no solo los datos del intestino humano, sino también una amplia gama de estudios relacionados con microbiomas en diferentes biomas, lo cual enriquece el análisis con una mayor diversidad de fuentes.

La decisión de realizar el análisis utilizando MGnify en lugar de HumGut se justifica por la capacidad de MGnify de ofrecer acceso más amplio a diversas muestras permiten una caracterización más completa del microbioma humano, facilitando la comparación entre diferentes estudios.

4.1.2 Protocolo de extracción de datos

- Nombre de la Base de Datos
- Descripción de la Base de Datos
- Enlace a la Base de Datos
- Número de Muestras Disponibles
- Tipos de Técnicas de Secuenciación Utilizadas
- Enfermedad Específica Estudiada
- Frecuencia de Actualización
- Diversidad de las Poblaciones Estudiadas

Nombre de la Base de Datos	Descripción	Enlace	Número de Muestras	Técnicas de Secuenciación	Enfermedad	Frecuencia de Actualización	Diversidad de las Poblaciones
HPMCD (Human Pan-Microbial Communities Database)	Agrupa comunidades microbianas humanas de diferentes estudios y proporciona datos metagenómicos de alta calidad.	http://baymax.erc.monash.edu/index.php	1,830 muestras	Metagenómica de genoma completo, Secuenciación de 16S rRNA	Enfermedad de Crohn, Colitis ulcerosa.	Anual	Alta
GMrepo (Gut Microbiota Repository)	Repositorio especializado en datos de microbiota intestinal, proporcionando información detallada sobre la composición y funcionalidad de los microbiomas intestinales	https://gmrepo.humangut.info/home	58,903 muestras	Secuenciación de 16S rRNA, Metagenómica de genoma completo	Enfermedad de Crohn, Colitis ulcerosa, Otras enfermedades inflamatorias intestinales	Anual	Alta
MGnify	Plataforma del EBI dedicada a la anotación, análisis y visualización de datos	https://www.ebi.ac.uk/metagenomics	Más de 100,000 muestras	Metagenómica de genoma completo, Metatranscriptómica,	Enfermedad de Crohn, Colitis ulcerosa, Otras	Trimestral	Alta
HumanMetagenomeDB (HMgDB)	Repositorio público de datos metagenómicos humanos, facilitando la búsqueda y selección de metagenomas para investigaciones biomédic	https://web.app.ufz.de/hmgdb/	69,822 metagenomas	Metagenómica de genoma completo, Secuenciación de 16S rRNA, Metatranscriptómica	Enfermedad de Crohn, Colitis ulcerosa, Otras enfermedades inflamatorias intestinales	Anual	Alta

Tabla 4. Protocolo de extracción de datos. Elaboración propia

4.2 Análisis descriptivo de las bases de datos identificadas

El análisis descriptivo de las bases de datos de secuenciación metagenómica se realizó utilizando Python, empleando las librerías pandas para la manipulación de datos y matplotlib para la visualización gráfica.

4.2.1 Distribución del número de muestras

El análisis de las bases de datos metagenómicas identificadas revela una disparidad significativa en la cantidad de muestras disponibles. MGnify, con 100,000 muestras lidera en cantidad y diversidad de biomas, lo que la convierte en una plataforma ideal para estudios comparativos y de seguimiento (Almeida et al., 2020; Richardson et al., 2023). Le sigue HumanMetagenomeDB con 69,822 muestras, que destaca por su riguroso enfoque en la curación de metadatos, permitiendo estudios reproducibles y comparables en el campo de la medicina personalizada (Kasmanas et al., 2021). GMrepoV2, con 58,903 muestras, se especializa en el microbioma intestinal, lo que facilita el análisis de marcadores microbianos en enfermedades como el cáncer colorrectal y la EC (Dai et al., 2022). Finalmente, HPMCD, con 1,830 muestras, aunque más pequeña, ofrece datos precisos y profundos sobre comunidades microbianas específicas, útiles para estudios enfocados en la composición microbiana (Chen et al., 2020).

4.2.2 Frecuencia de actualización

La frecuencia de actualización de las bases de datos es clave para mantener la relevancia en estudios microbiológicos. MGnify se actualiza trimestralmente, lo que la hace ideal para estudios longitudinales (Richardson et al., 2023). HumanMetagenomeDB y GMrepo, con actualizaciones anuales, aseguran la incorporación de nuevos estudios y la estandarización necesaria para investigaciones comparativas (Kasmanas et al., 2021; Dai et al., 2022). HPMCD también se actualiza anualmente, manteniendo datos consistentes y de alta calidad en estudios específicos sobre microbiomas humanos (Chen et al., 2020). Estas frecuencias de actualización reflejan el compromiso con la calidad y accesibilidad de los datos, fundamentales para avanzar en el estudio de las EII.

4.2.3 Diversidad de las poblaciones

Las bases de datos presentan una alta diversidad de poblaciones, un aspecto clave para estudiar la complejidad del microbioma en diversos contextos geográficos y étnicos. MGnify se destaca por la representación global de muestras, facilitando estudios comparativos entre biomas y etnias (Richardson et al., 2023). HumanMetagenomeDB incluye datos demográficos detallados, que son cruciales para estudios de EC y CU (Kasmanas et al., 2021). GMrepo y

HPMCD, aunque más específicos, aportan datos valiosos sobre poblaciones controladas en condiciones patológicas, siendo herramientas esenciales en estudios especializados (Dai et al., 2022; Chen et al., 2020).

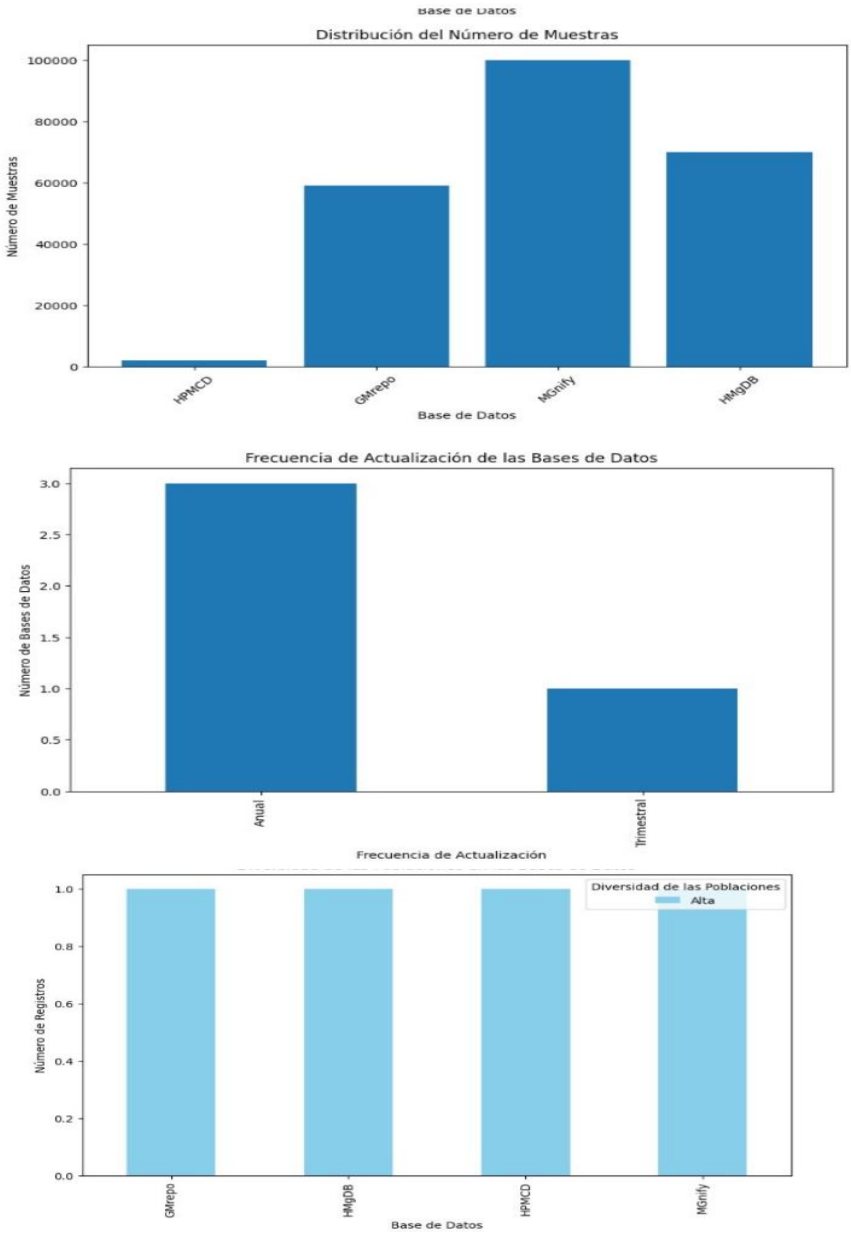


Figura 3. Análisis descriptivo de las bases de datos identificadas. Creado en Python librería Matplotlib

A continuación, se presentará un análisis descriptivo de los datos de secuenciación metagenómica en las bases de datos HPMCD (Human Pan-Microbial Communities Database), GMrepo (Gut Microbiota Repository), MGnify, HumanMetagenomeDB (HMgDB) enfocadas en la distribución de variables clave como estado de la enfermedad, género, etnicidad, nivel taxonómico, bioma, centro de investigación y métricas de calidad de secuenciación.

4.2.4 HPMCD

La base de datos HPMCD se destaca por su enfoque en la curación detallada de comunidades microbianas humanas provenientes de estudios metagenómicos. Proporciona un conjunto de datos revisados que facilita la comparación entre diferentes comunidades microbianas y sus roles tanto en condiciones normales como patológicas (Chen et al., 2020). Es particularmente útil para investigar cómo las diferencias en la microbiota influyen en la respuesta a EII como la EC y la CU (Chen et al., 2020). Aunque HPMCD tiene una cobertura de poblaciones específica, esta limitación se ve compensada por la alta calidad y relevancia de los datos, consolidándose como una herramienta esencial para la investigación avanzada en microbiota y salud humana.

4.2.4.1 Distribución del Estado de la Enfermedad

La mayoría de las muestras en HPMCD están asociadas a la EC, con un total de 24 muestras, en comparación con 14 muestras de colitis ulcerosa (CU). Este desequilibrio refleja una mayor representación de EC, lo que puede sesgar los análisis hacia esta enfermedad, limitando la comprensión detallada de la microbiota en la CU (Chen et al., 2020).

4.2.4.2 Distribución de Género

La distribución de género en HPMCD muestra una marcada predominancia femenina, con 30 muestras de mujeres frente a solo 8 de hombres. Este desbalance de género podría influir en la interpretación de los datos y los resultados de los análisis, ya que refleja una representación desigual entre los sexos (Chen et al., 2020).

4.2.4.3 Distribución de Etnicidad

Las muestras en HPMCD provienen exclusivamente de sujetos de origen europeo, lo que limita la diversidad étnica en los datos. Esta falta de diversidad podría restringir la

generalización de los resultados y su aplicabilidad a otras poblaciones, afectando la comprensión de cómo la variabilidad étnica influye en la microbiota en el contexto de enfermedades inflamatorias intestinales (Chen et al., 2020).

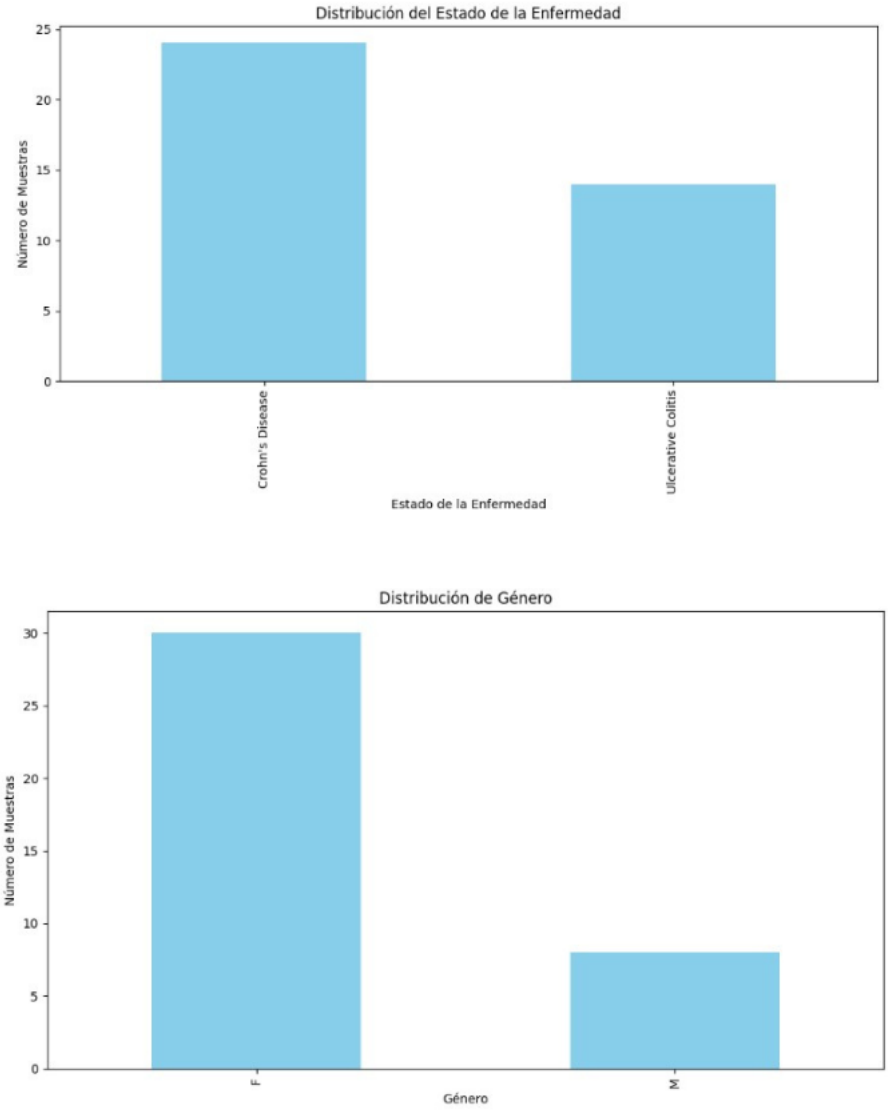


Figura 4. HPMCD (Human Pan-Microbial Communities Database)

4.2.5 GMrepoV2

La base de datos GMrepoV2 se distingue por su rigurosidad en la curación manual de datos y la aplicación de estándares consistentes en la anotación de secuencias, lo cual mejora

la calidad y fiabilidad de los resultados obtenidos (Dai et al., 2022). Su diseño facilita la integración y comparación de datos provenientes de múltiples estudios, lo que la convierte en una herramienta valiosa para comprender las complejas interacciones microbianas en el intestino humano, particularmente en relación con enfermedades como la EC y la EII. En cuanto a EC, los análisis de la distribución del nivel taxonómico revelan una notable variabilidad, lo que refleja la complejidad del microbioma intestinal en esta enfermedad. Además, las especies microbianas más predominantes incluyen Barnesiella intestinihominis y Bacteroides finegoldii, lo que subraya su posible papel en la microbiota alterada en la EC (Dai et al., 2022).

Para la EII, los datos de GMrepoV2 muestran una concentración notable de muestras bajo un único fenotipo específico, lo que sugiere un enfoque uniforme en la clasificación de los datos a nivel de especie. La distribución del nivel taxonómico también es diversa, destacándose el nivel 1681 como el más frecuente, seguido por otros niveles menos representados, lo que subraya la implicación de múltiples taxones en la fisiopatología de la EII. Entre las especies microbianas más frecuentes en las muestras de EII se encuentra Gemmiger formicilis, presente en más de 30 muestras, lo que podría ser relevante para comprender su papel en los procesos inflamatorios de la enfermedad (Dai et al., 2022).

Cabe mencionar que las figuras relacionadas con la distribución taxonómica EC y la EII, que proporcionan detalles adicionales sobre la diversidad microbiana a diferentes niveles taxonómicos, se encuentran en el Anexo 1. Estas gráficas complementan el análisis principal al ofrecer una visión más detallada de la clasificación microbiana y las interacciones en estas enfermedades (Dai et al., 2022).

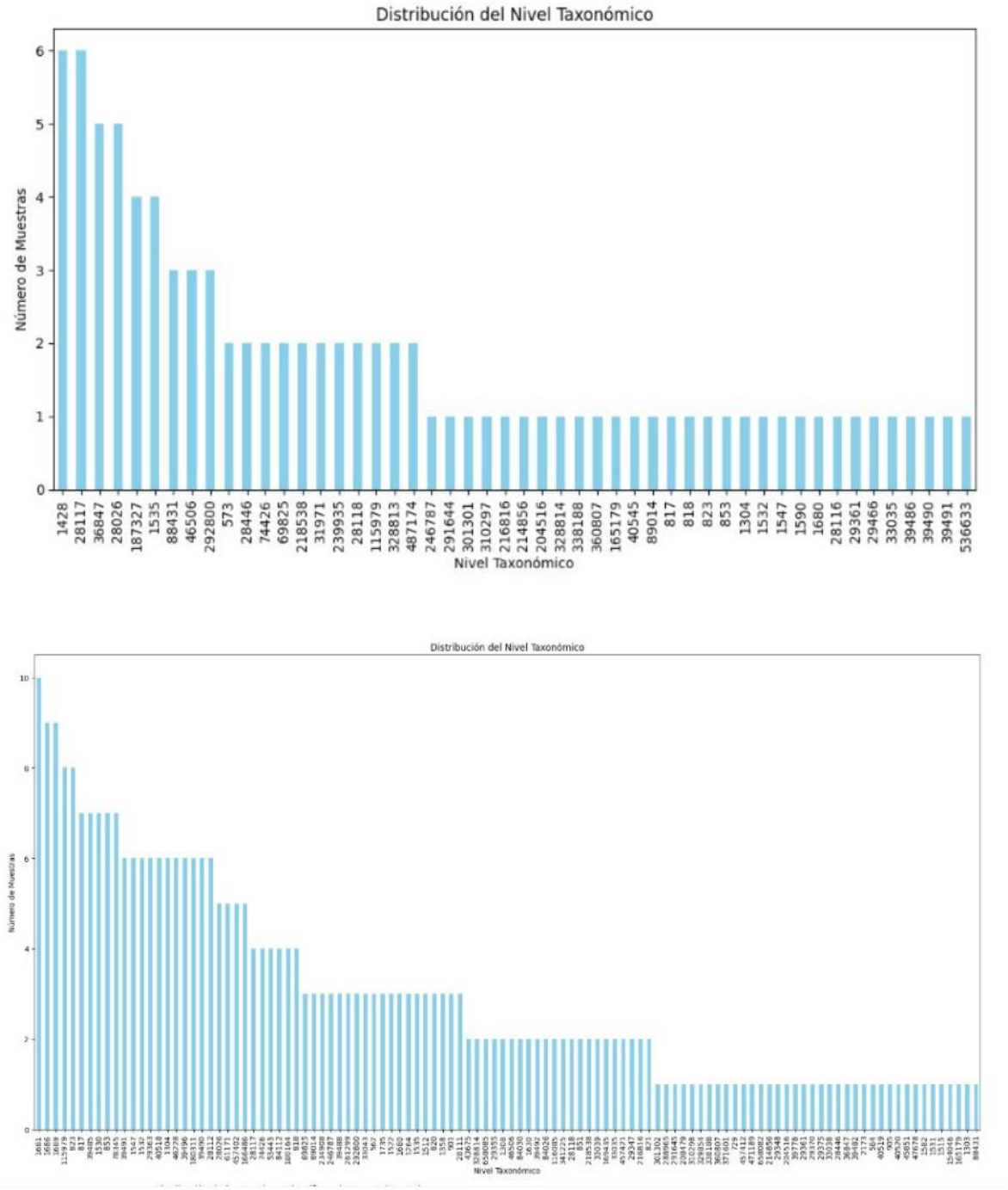


Figura 5. Distribución del nivel taxonómico. Creado en Python librería Matplotlib

4.2.6 MGnify

La base de datos MGnify, gestionada por el EMBL-EBI, se presenta como una plataforma robusta para el análisis y comparación de datos metagenómicos, especialmente útil para el estudio del microbioma intestinal en diversas condiciones de salud, incluyendo las EII (Richardson et al., 2023). Ofrece un entorno colaborativo que permite a los investigadores acceder a datos sobre la diversidad microbiana en múltiples entornos, facilitando un análisis detallado de la relación entre la microbiota y EC o la CU.

En relación con la EC, la mayoría de las muestras en MGnify provienen del bioma fecal, lo que subraya la importancia de esta área en la investigación de esta patología. Además, los centros de investigación que contribuyen con datos provienen de una amplia diversidad geográfica, resaltando el esfuerzo colaborativo global en el estudio de la EC. Los proyectos asociados cubren una gran variedad de temas, desde el análisis de la variabilidad temporal de la microbiota hasta la correlación de la microbiota con enfermedades crónicas, demostrando la riqueza de datos que MGnify ofrece para avanzar en la investigación de la EC (Richardson et al., 2023).

En cuanto a la CU, se observa una mayor diversidad en los biomas representados, abarcando no solo el bioma fecal, sino también otros tejidos gastrointestinales como el colon sigmoideo y el intestino grueso. La procedencia de los datos también refleja un esfuerzo global, con contribuciones de centros de investigación de renombre internacional como Harvard y el MIT. Además, los proyectos relacionados con CU incluyen una variedad de enfoques, destacándose el interés en el trasplante de microbiota fecal como una intervención terapéutica en esta enfermedad, Anexo 1 (Richardson et al., 2023).

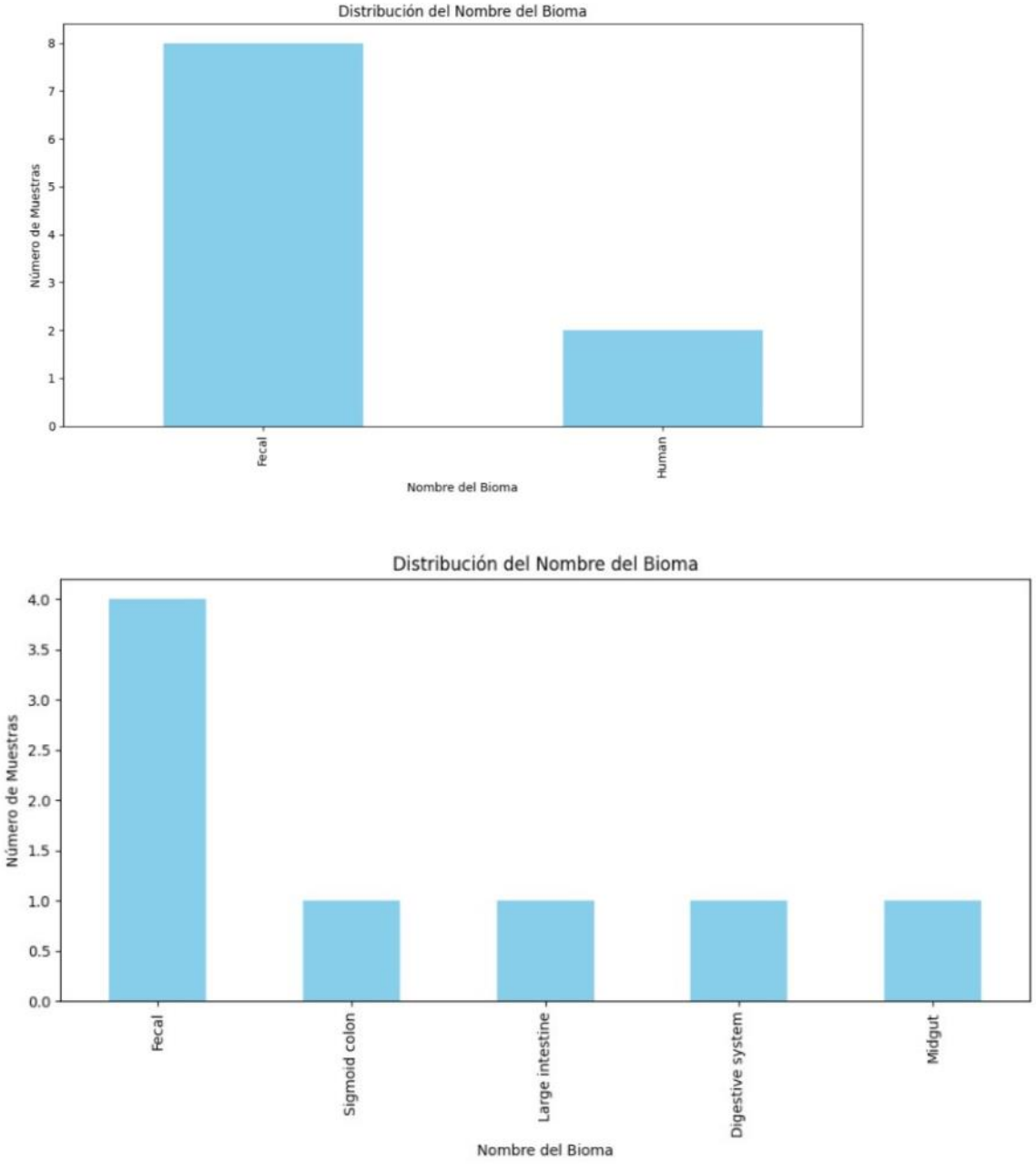


Figura 6. Distribución del nombre del bioma MGnify. Creado con Python libería Matplotlib

4.2.7 HMgDB

HMgDB se destaca como un recurso fundamental para el análisis y comparación de datos metagenómicos de origen humano, específicamente en el estudio de EII como la EC y la colitis ulcerativa CU. Su enfoque en la curación y estandarización de los metadatos asegura una alta calidad de los datos y permite a los investigadores explorar el microbiota a través de diferentes cohortes y estudios (Kasmanas et al., 2021). La EC es una de las patologías más representadas en HMgDB, con un total de 679 muestras, lo que permite una investigación exhaustiva de los patrones microbianos asociados a esta enfermedad y su diferenciación con otras patologías inflamatorias (Kasmanas et al., 2021).

La plataforma de secuenciación utilizada predominantemente en HMgDB para las muestras de EC y CU es Illumina, lo que garantiza una consistencia en los datos generados y minimiza las variaciones técnicas. Esta uniformidad en la tecnología empleada es esencial para asegurar que los análisis comparativos entre estudios sean precisos y confiables (Kasmanas et al., 2021). En cuanto a la calidad de las secuencias, la base de datos destaca por su alto estándar, con un valor Q30 superior al 99.9% en la mayoría de las lecturas, asegurando así una alta precisión en la identificación y cuantificación de las especies microbianas en ambas enfermedades (Kasmanas et al., 2021).

En relación con la CU, HMgDB incluye un total de 373 muestras, representando una proporción significativa en los estudios de EII. La consistencia en la calidad de las muestras secuenciadas con Illumina refuerza la fiabilidad de los datos, lo que es esencial para garantizar que los análisis metagenómicos sean precisos y que los resultados de los estudios sean comparables entre distintas investigaciones, Anexo 1 (Kasmanas et al., 2021).

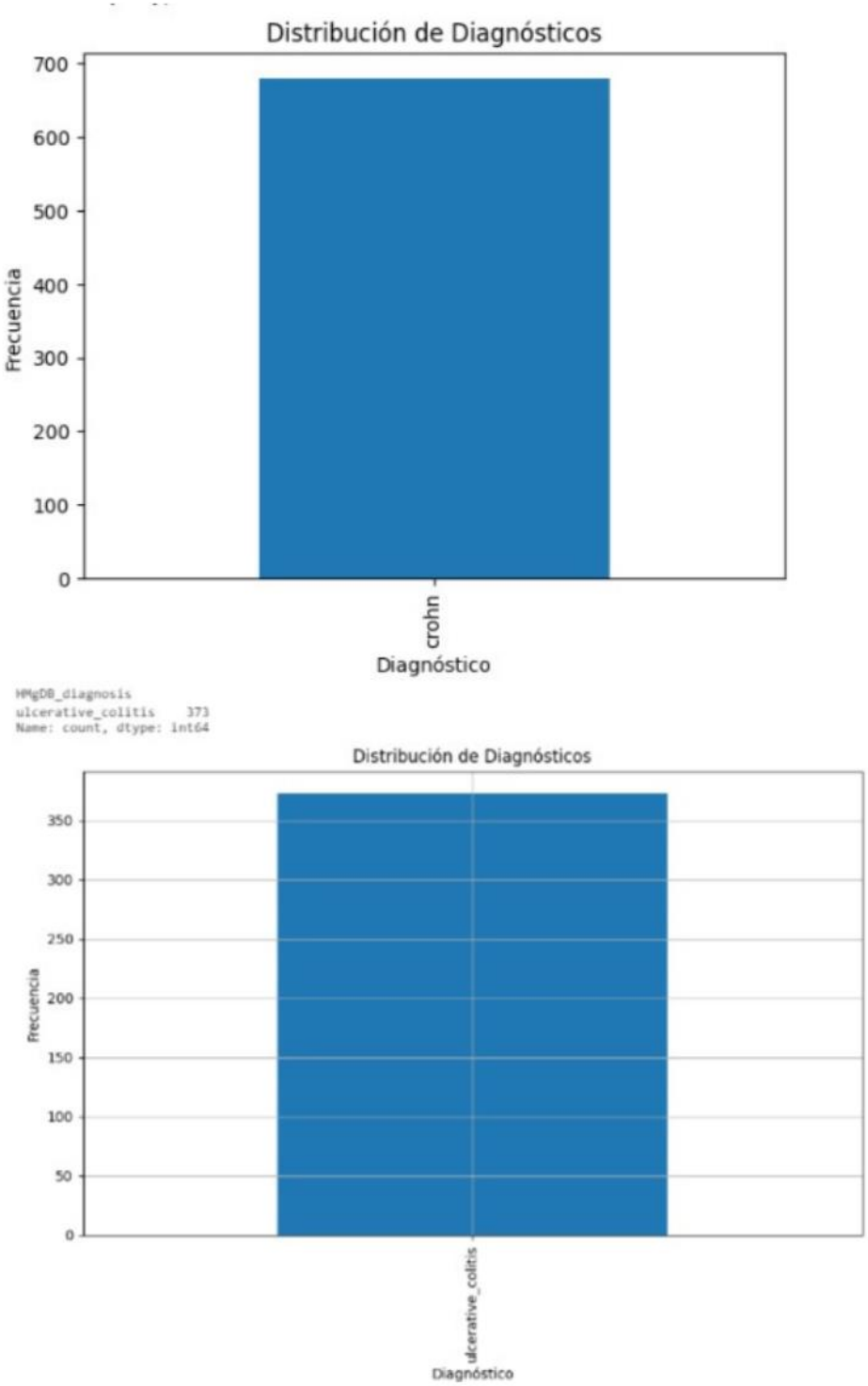


Figura 7. Distribución de diagnóstico HMgDB. Creado con Python libería Matplotlib

4.3 Análisis de adherencia a los principios FAIR en bases de datos metagenómicas

La evaluación de la adherencia a los principios FAIR en las bases de datos metagenómicas (GMrepo v2, HPMCD, MGnify y HumanMetagenomeDB) reveló una sólida implementación de estos principios, con especial énfasis en la interoperabilidad y accesibilidad de los datos. Todas las bases de datos presentaron identificadores únicos y persistentes (DOI, ARK, Handle), lo que asegura la trazabilidad y reconocimiento de los datos en la comunidad científica. Además, los metadatos en todas las bases de datos incluían referencias claras a estos identificadores, lo que facilita la búsqueda y reutilización de la información.

El uso de esquemas de metadatos legibles por máquina fue consistente, permitiendo la integración eficiente de los datos con otros conjuntos, promoviendo la interoperabilidad. Asimismo, la mayoría de los datos se encuentran en repositorios especializados y de acceso público, lo que refuerza tanto la especificidad como la accesibilidad a la información. Los métodos de acceso a los datos incluyen desde descargas directas hasta APIs estándar, lo que facilita su reutilización en diferentes entornos de investigación.

Por último, todas las bases de datos utilizaron formatos de datos abiertos, lo que asegura un acceso universal sin depender de software propietario. Se destacó el uso de vocabularios estandarizados y enlaces URI en los metadatos, lo que mejora la comprensión y navegación entre conjuntos de datos relacionados. Además, las bases de datos emplearon licencias estándar como Creative Commons, garantizando condiciones claras para la reutilización de los datos y promoviendo el intercambio ético de información.

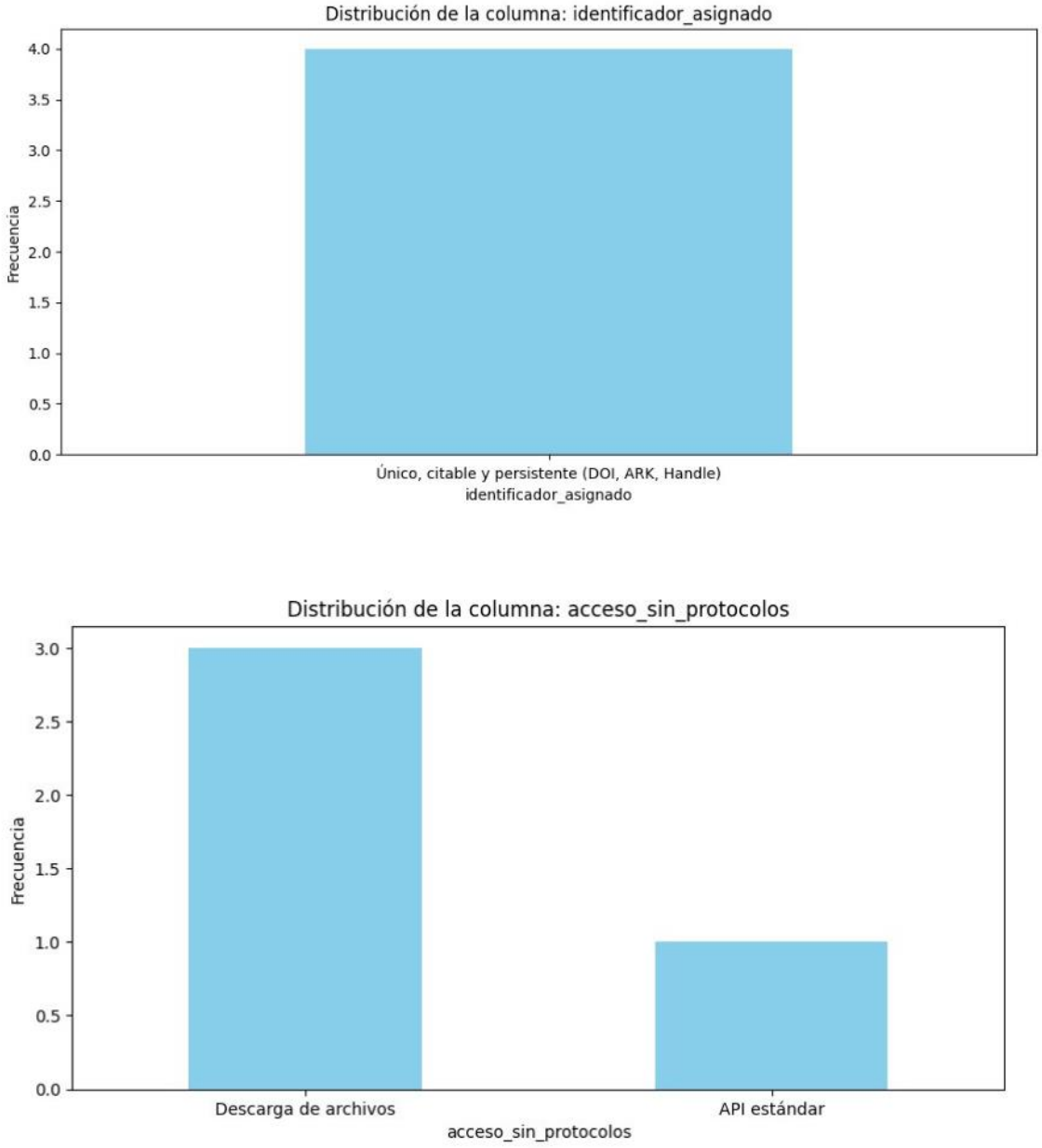


Figura 8. Análisis de adherencia a los principios FAIR en bases de datos metagenómicas. Creado con Python libería Matplotlib

5 Discusión

Los resultados obtenidos revelan un panorama complejo en la adherencia de las bases de datos metagenómicas a los principios FAIR, resaltando tanto avances como desafíos críticos. Al evaluar bases de datos como GMrepo v2, HPMCD, MGnify y HMgDB, se encontró que todas muestran un compromiso significativo con la accesibilidad y la interoperabilidad mediante el uso de identificadores persistentes y esquemas de metadatos legibles por máquina. Estos hallazgos confirman parcialmente la hipótesis planteada, evidenciando que, aunque las bases de datos están accesibles y sus datos son en su mayoría interoperables, las inconsistencias en la estandarización de metadatos y la falta de ontologías comunes limitan su reutilización efectiva en modelos de aprendizaje automático. Esto subraya que, a pesar del esfuerzo por cumplir con los principios FAIR, los repositorios aún enfrentan barreras significativas que impactan su plena utilidad en aplicaciones ML.

Sin embargo, el análisis detallado mostró que la implementación de protocolos de acceso y la adopción de vocabularios normalizados presentan variabilidad considerable, lo que confirma la presencia de brechas en la estandarización y reutilización de los datos. Aunque se observó una alineación con algunos principios FAIR, la falta de uniformidad en los metadatos y la diversidad en la implementación de ontologías crean un escenario fragmentado, donde los datos, aunque técnicamente accesibles, no logran integrarse de manera fluida entre diferentes estudios y plataformas. Esto refuerza la hipótesis de que, si bien existen esfuerzos notables hacia la mejora de la interoperabilidad y accesibilidad, persisten desafíos estructurales y técnicos que limitan la verdadera reutilización de los datos, confirmando la necesidad de una mayor estandarización para maximizar el valor y el impacto de estas bases de datos en la investigación y en el desarrollo de modelos predictivos en EII.

Los hallazgos de este trabajo de fin de master reflejan un escenario donde la adherencia a los principios FAIR se presenta como un objetivo parcialmente alcanzado, con ecos notables en la literatura, pero también con divergencias significativas que subrayan los retos persistentes. Por ejemplo, el 1000IBD Project, al integrar datos multi-ómicos de más de 1000 pacientes con EII, demuestra un avance crucial hacia la interoperabilidad y accesibilidad de los datos, permitiendo que la investigación traslacional avance hacia la medicina de precisión (Imhann et al., 2019). Sin embargo, esta gran promesa contrasta con las dificultades

en la estandarización de los metadatos y la falta de ontologías abiertas, una problemática que también emerge en nuestros resultados, sugiriendo un paisaje científico aún fragmentado donde la reutilización efectiva se ve obstaculizada.

En concordancia, Jin et al. (2024) revelan que la falta de estandarización y la curación inconsistente de los datos metagenómicos impactan profundamente la capacidad de integrar y comparar estudios sobre EC y CU. Similar a lo observado en nuestro análisis, estos estudios destacan que la interoperabilidad sigue siendo parcial y frecuentemente deficiente, limitando la utilidad de los datos en aplicaciones de ML.

El Inflammatory Bowel Diseases Integrated Resources Portal (IBDIRP), presentado por Nie et al. (2024), refuerza la importancia de un enfoque meticuloso hacia la curación manual de los datos, consolidando la adherencia a los principios FAIR en el ámbito de las EII. Este modelo de base de datos integrada se posiciona como un ejemplo de cómo la estandarización y la accesibilidad pueden mejorar la calidad y la reutilización de los datos. No obstante, como también se refleja en nuestros resultados, la implementación de ontologías comunes aún se percibe insuficiente, limitando la capacidad de conectar e interpretar los datos a través de estudios y plataformas diferentes.

Además, el estudio de Van der Lelie et al. (2021) sobre consorcios bacterianos diseñados racionalmente expone cómo la integración efectiva de datos multi-ómicos puede guiar el diseño de terapias innovadoras, enfocadas en restaurar la homeostasis intestinal. Este enfoque, alineado con los principios de interoperabilidad y reutilización, subraya la necesidad de bases de datos altamente caracterizadas y estandarizadas. Sin embargo, la falta de uniformidad en los protocolos y la variabilidad en la calidad de los datos siguen siendo barreras significativas, tal como se evidencia en nuestro trabajo, dificultando la traducción de los datos en aplicaciones clínicas efectivas.

Al aplicar un marco metodológico riguroso y detallado, nuestro trabajo no solo evalúa la alineación de las bases de datos con los principios FAIR, sino que también expone las brechas que otros enfoques podrían pasar por alto. La revisión bibliografica, guiada por PRISMA, asegura una evaluación crítica de la literatura relevante, proporcionando una base para comparar y contextualizar estos datos dentro del panorama actual de la bioinformática aplicada a las EII. Así, nuestro enfoque no solo documenta las deficiencias existentes, sino que

también actúa como una guía estratégica para futuras mejoras en la estandarización y en la optimización de los datos para aplicaciones clínicas y de investigación, fortaleciendo la capacidad de traducir estos avances en soluciones terapéuticas efectivas.

Las principales limitaciones de este trabajo se centran en el sesgo de selección de bases de datos y la falta de diversidad étnica en las muestras analizadas. Al seleccionar repositorios que cumplen con criterios mínimos de accesibilidad y adherencia a los principios FAIR, es posible que se hayan excluido otras bases de datos que, aunque menos accesibles o estandarizadas, podrían ofrecer perspectivas adicionales sobre la variabilidad y calidad de los datos metagenómicos. Este sesgo en la selección limita la capacidad de los resultados para capturar la complejidad completa del panorama de datos disponible, restringiendo la aplicabilidad y comparación entre diferentes estudios.

Además, la falta de diversidad étnica en las muestras representa una limitación significativa, ya que la mayoría de los datos provienen de sujetos de origen europeo, lo cual reduce la representatividad y la generalizabilidad de los hallazgos a poblaciones más diversas. Esta falta de inclusión puede sesgar la interpretación de los datos y las conclusiones derivadas, afectando la capacidad de los modelos predictivos y terapéuticos para ser efectivos en contextos clínicos globales.

En resumen, mientras que la literatura muestra progresos notables hacia la adherencia a los principios FAIR, especialmente en términos de accesibilidad y accesibilidad parcial, persisten desafíos críticos en la estandarización y en la implementación de protocolos comunes. Estos obstáculos, reflejados tanto en nuestros resultados como en estudios previos, evidencian la necesidad de esfuerzos coordinados para superar las barreras y maximizar el valor de los datos metagenómicos en la investigación de las EII.

6 Conclusiones

Este trabajo evaluó la utilidad de los conjuntos de datos de secuenciación metagenómica relacionados con EC, CU y EII para su aplicación en modelos de aprendizaje automático, enfocándose en la integridad de los datos y su adherencia a los principios FAIR. Se concluye que, aunque las bases de datos revisadas presentan un compromiso considerable con los principios de accesibilidad e interoperabilidad, persisten limitaciones significativas en

términos de estandarización y reutilización. Estas barreras dificultan la integración de los datos en modelos de ML y aplicaciones clínicas.

Se identificaron y clasificaron cuatro bases de datos abiertas (GMrepo v2, HPMCD, MGnify y HumanMetagenomeDB) que contienen datos de secuenciación metagenómica relacionados con EC, CU y otras EII. Estas bases de datos proporcionan volúmenes sustanciales de datos y características variadas que permiten una investigación detallada. Sin embargo, la falta de uniformidad en la curación y estandarización de los metadatos afecta su comparabilidad, lo que podría limitar su potencial en estudios más amplios y modelos predictivos.

El análisis estadístico de los datos metagenómicos mostró disparidades considerables en las variables clave como el estado de la enfermedad, género, etnicidad, nivel taxonómico y bioma. Se detectó una falta de diversidad étnica, ya que la mayoría de los datos provienen de sujetos de origen europeo, lo que limita la aplicabilidad de los hallazgos a nivel global. Sin embargo, la homogeneidad en el uso de la plataforma de secuenciación Illumina asegura una mayor consistencia en los análisis.

El análisis de adherencia a los principios FAIR realizado mediante el formulario creado por la Universidad Europea de Madrid, mostró que las bases de datos cumplen en gran medida con los principios de accesibilidad e interoperabilidad, gracias a la implementación de identificadores únicos y vocabularios estandarizados. No obstante, se encontraron deficiencias en la estandarización y reutilización de los datos, lo que limita su potencial para integrarse en modelos de ML y reducir las barreras para su reutilización en otros estudios.

7 Bibliografía

Abraham, C., & Cho, J. H. (2009). Inflammatory bowel disease. *New England Journal of Medicine*, *361(21), 2066–2078.* https://doi.org/10.1056/NEJMra0804647

Achudhan, A. B., Kannan, P., & Saleena, L. M. (2024). A review of web-based metagenomics platforms for analyzing next-generation sequence data. *Biochemical Genetics, 62, 621–632*. https://doi.org/10.1007/s10528-023-10467-w

Ahmad, H. A., East, J. E., Panaccione, R., Travis, S., Canavan, J. B., Usiskin, K., & Byrne, M. F. (2023). Artificial intelligence in inflammatory bowel disease: Implications for clinical practice and future directions. *Intestinal Research, 21(3), 283–294*. https://doi.org/10.5217/ir.2023.00020

Almeida, A., Mitchell, A. L., Boland, M., Forster, S. C., Gloor, G. B., Tarkowska, A., Lawley, T. D., & Finn, R. D. (2020). HumGut: A comprehensive human gut prokaryotic genomes collection filtered by metagenome data. *Nature Biotechnology, 38(10), 1139–1143*. https://doi.org/10.1038/s41587-020-0603-3

Chen, Y., Li, Z., Li, Z., Qi, Y., & Liu, Z. (2020). HPMCD: The database of human microbial communities from metagenomic datasets. *Journal of Genomics, 8, 39–45*. https://doi.org/10.7150/jgen.46654

Cleo, G., Scott, A. M., Islam, F., Julien, B., & Beller, E. (2019). Usability and acceptability of four systematic review automation software packages: A mixed method design. *Systematic Reviews, 8(1), Article 145*. https://doi.org/10.1186/s13643-019-1069-6

Crohn's & Colitis Foundation of America. (2014). *Inflammatory bowel diseases fact book*. Crohn's & Colitis Foundation of America. http://www.ccfa.org

Dai, D., Zhu, J., Sun, C., Li, M., Liu, J., Wu, S., Ning, K., He, L.-J., Zhao, X.-M., & Chen, W.-H. (2022). GMrepo v2: A curated human gut microbiome database with special focus on disease markers and cross-dataset comparison. *Nucleic Acids Research, 50(D1)*, D777–D784. https://doi.org/10.1093/nar/gkab1019

Fiocchi, C. (2023). Omics and multi-omics in IBD: No integration, no breakthroughs. *International Journal of Molecular Sciences, 24(19)*, Article 14912. https://doi.org/10.3390/ijms241914912

Freeman, K., Ryan, R., Parsons, N., Taylor-Phillips, S., Willis, B. H., & Clarke, A. (2021). The incidence and prevalence of inflammatory bowel disease in UK primary care: A retrospective cohort study of the IQVIA Medical Research Database. *BMC Gastroenterology, 21, Article 139*. https://doi.org/10.1186/s12876-021-01716-6

Gonzalez, C. G., Mills, R. H., Zhu, Q., Sauceda, C., Knight, R., Dulai, P. S., & Gonzalez, D. J. (2022). Location-specific signatures of Crohn's disease at a multi-omics scale. *Microbiome, 10, Article 133*. https://doi.org/10.1186/s40168-022-01331-x

González Soltero, M. del R., Castro, A. B., & Pino, D. (2024). Autoevaluación de datos FAIR. *Proyecto creado en REDCap.* https://redcap.universidadeuropea.com/surveys/?s=L89WYPJ3RFL3D7WP

Hemmer, A., Forest, K., Rath, J., & Bowman, J. (2023). Inflammatory bowel disease: A concise review. *South Dakota Medicine, 76(9), 416–423*.

Imhann, F., van der Velde, K. J., Barbieri, R., Alberts, R., Voskuil, M. D., Vich Vila, A., ... & Weersma, R. K. (2019). The 1000IBD project: Multi-omics data of 1000 inflammatory bowel disease patients; data release 1. *BMC Gastroenterology, 19(1), Article 5*. https://doi.org/10.1186/s12876-018-0917-5

Javaid, A., Shahab, O., Adorno, W., Fernandes, P., May, E., & Syed, S. (2021). Machine learning predictive outcomes modeling in inflammatory bowel diseases. *Inflammatory Bowel Diseases, 28(4), 819–829*. https://doi.org/10.1093/ibd/izab187

Jin, D.-M., Morton, J. T., & Bonneau, R. (2024). Meta-analysis of the human gut microbiome uncovers shared and distinct microbial signatures between diseases. *mSystems*. https://doi.org/10.1128/msystems.00295-24

Nie, K., Cai, Q., Ma, K., Luo, W., Wu, X., Chen, X., Cai, L., Deng, M., Yang, Y., & Wang, X. (2024). An inflammatory bowel diseases integrated resources portal (IBDIRP). *Database, 2024, baad097*. https://doi.org/10.1093/database/baad097

Kang, D.-Y., Park, J.-L., Yeo, M.-K., Kang, S.-B., Kim, J.-M., Kim, J. S., & Kim, S.-Y. (2023). Diagnosis of Crohn's disease and ulcerative colitis using the microbiome. *BMC Microbiology, 23, Article 336*. https://doi.org/10.1186/s12866-023-03084-5

Kasmanas, J. C., Bartholomäus, A., Corrêa, F. B., Tal, T., Jehmlich, N., Herberth, G., & Rocha, U. N. (2021). HumanMetagenomeDB: A public repository of curated and standardized metadata for human metagenomes. *Nucleic Acids Research, 49(D1), D743–D750*. https://doi.org/10.1093/nar/gkaa1031

Krassowski, M., Das, V., Sahu, S. K., & Misra, B. B. (2020). State of the field in multi-omics research: From computational needs to data mining and sharing. *Frontiers in Genetics, 11, Article 610798*. https://doi.org/10.3389/fgene.2020.610798

Manski, S., Noverati, N., Policarpo, T., Rubin, E., & Shivashankar, R. (2024). Diet and nutrition in inflammatory bowel disease: A review of the literature. *Crohn's & Colitis 360, 6(1),* 1–10. https://doi.org/10.1093/crocol/otad077

Nie, K., Cai, Q., Ma, K., Luo, W., Wu, X., Chen, X., Cai, L., Deng, M., Yang, Y., & Wang, X. (2024). An inflammatory bowel diseases integrated resources portal (IBDIRP). *Database, 2024, baad097*. https://doi.org/10.1093/database/baad097

Ning, L., Zhou, Y.-L., Sun, H., Zhang, Y., Shen, C., Wang, Z., Xuan, B., Zhao, Y., Ma, Y., Yan, Y., Tong, T., Huang, X., Hu, M., Zhu, X., Ding, J., Zhang, Y., Cui, Z., Fang, J.-Y., Chen, H., & Hong, J. (2023). Microbiome and metabolome features in inflammatory bowel disease via multi-omics integration analyses across cohorts. *Nature Communications, 14, Article 7135*. https://doi.org/10.1038/s41467-023-42788-0

Preto, A. J., Chanana, S., Ence, D., Healy, M. D., Domingo-Fernández, D., & West, K. A. (2024). Multi-omics data integration identifies novel biomarkers and patient subgroups in inflammatory bowel disease. *medRxiv*. https://doi.org/10.1101/2024.07.23.24310846

Richardson, L. J., Allen, B., Baldi, G., Beracochea, M., Bileschi, M., Burdett, T., Burgin, J., Caballero-Pérez, J., Cochrane, G., Colwell, L., Curtis, T., Escobar-Zepeda, A., Gurbich, T., Kale, V., Korobeynikov, A., Raj, S., Rogers, A. B., Sakharova, E., ... & Finn, R. D.

(2023). MGnify: The microbiome sequence data analysis resource in 2023. *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkac1080

Singular Genomics. (2023). What are quality scores? Singular Knowledge Base. *https://knowledge.singulargenomics.com/g4-ug-quality-scores.htm*

Schmieder, R., & Edwards, R. (2011). Quality control and preprocessing of metagenomic datasets. *Bioinformatics, 27(6), 863–864*. https://doi.org/10.1093/bioinformatics/btr026

Seyed Tabib, N. S., Madgwick, M., Sudhakar, P., Verstockt, B., Korcsmaros, T., & Vermeire, S. (2020). Big data in IBD: Big progress for clinical practice. *Gut, 69(8), 1520–1532*. https://doi.org/10.1136/gutjnl-2019-320065

Shan, Y., Lee, M., & Chang, E. B. (2022). The gut microbiome and inflammatory bowel diseases. *Annual Review of Medicine, 73, 455–468*. https://doi.org/10.1146/annurev-med-042320-021020

Syed, A. H., Abujabal, H. A. S., Ahmad, S., Malebary, S. J., & Alromema, N. (2024). Advances in inflammatory bowel disease diagnostics: Machine learning and genomic profiling reveal key biomarkers for early detection. *Diagnostics, 14(1182), 1–16*. https://doi.org/10.3390/diagnostics14111182

Valles-Colomer, M., Darzi, Y., Vieira-Silva, S., Falony, G., & Raes, J. (2016). Meta-omics in inflammatory bowel disease: Applications, challenges, and guidelines. *Journal of Crohn's and Colitis*, 10(6), 735–746. https://doi.org/10.1093/ecco-jcc/jjw024

Van der Lelie, D., Oka, A., Taghavi, S., Umeno, J., ... et al. (2021). Rationally designed bacterial consortia to treat chronic immune-mediated colitis and restore intestinal

homeostasis. *Nature Communications, 12, Article 3105*.

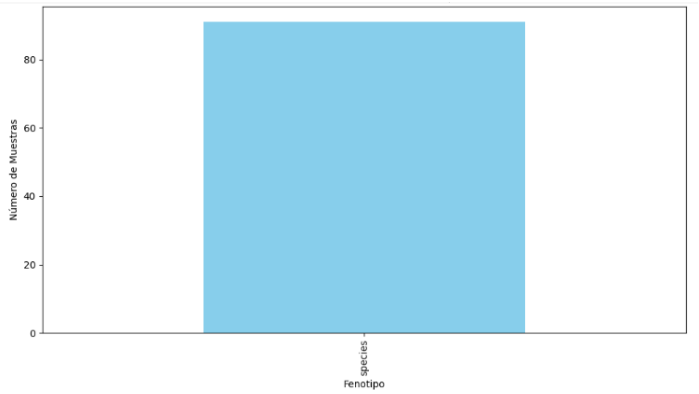
https://doi.org/10.1038/s41467-021-23481-8

Veauthier, B., & Hornecker, J. R. (2018). Crohn's disease: Diagnosis and management. *American Family Physician, 98(11), 661–669*.

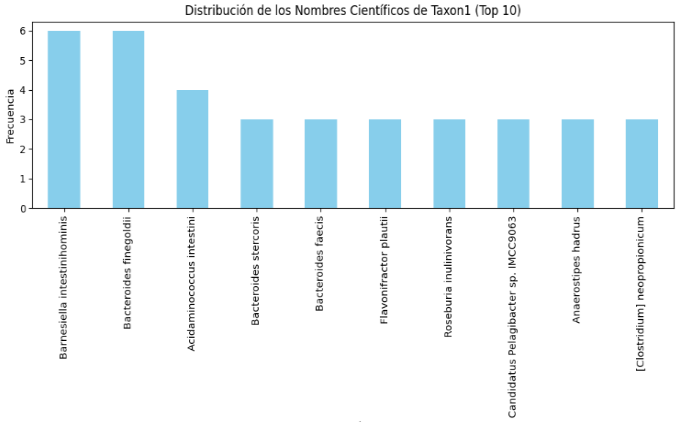
Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., & Mons, B. (2016). The FAIR guiding principles for scientific data management and stewardship. *Scientific Data, 3, Article 160018*.https://doi.org/10.1038/sdata.2016.18

8 Anexo 1

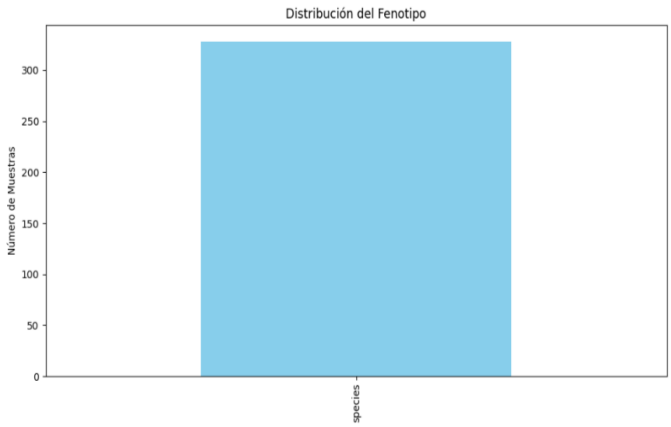
Se presentan las figuras de los resultados del análisis estadístico para las diferentes bases de datos.



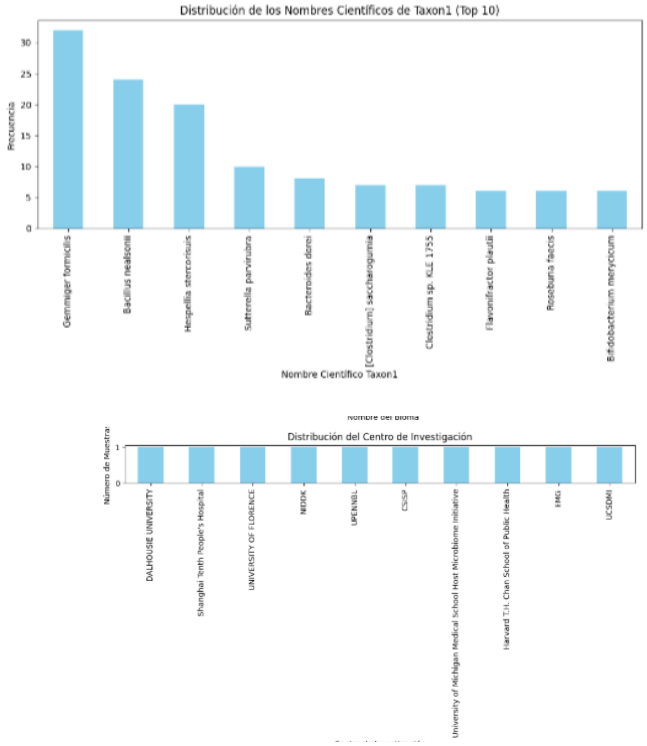
Distribución del Fenotipo en EC GMrepoV2. Elaboración propia con librería matplotlib. Python



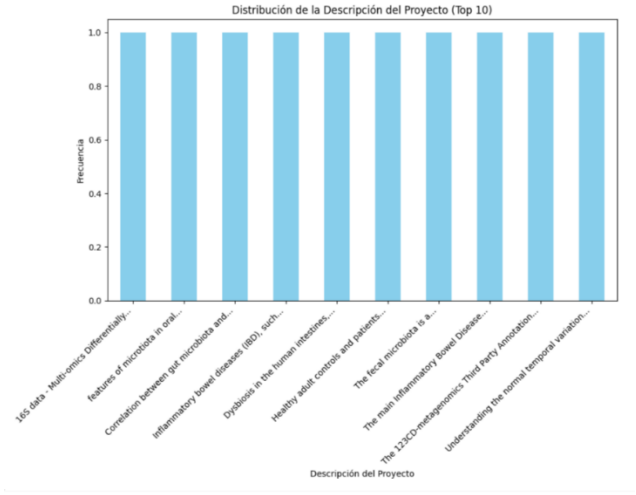
Distribución de los nombres científicos de Taxon 1 (Top 10) para la EC GMrepoV2. Elaboración propia con librería matplotlib. Python



Distribución del Fenotipo en la EII GMrepoV2. Elaboración propia con librería matplotlib. Python

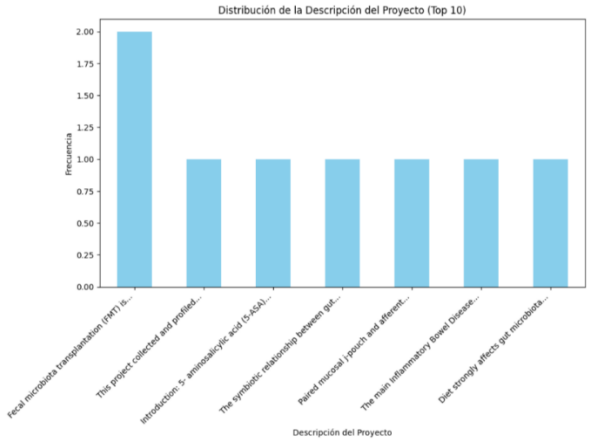


Distribución del centro de investigación para EC MGnify. Elaboración propia con librería matplotlib. Python

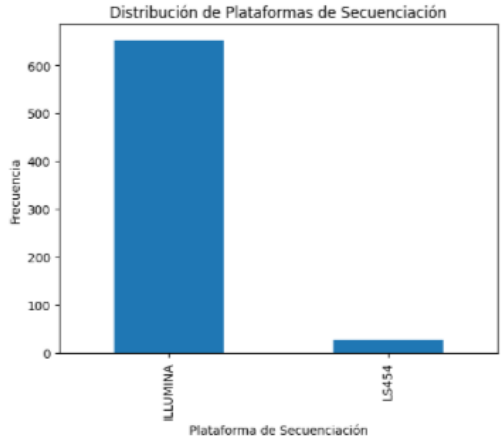


Distribución de la descripción del proyecto para EC MGnify. Elaboración propia con librería matplotlib. Python

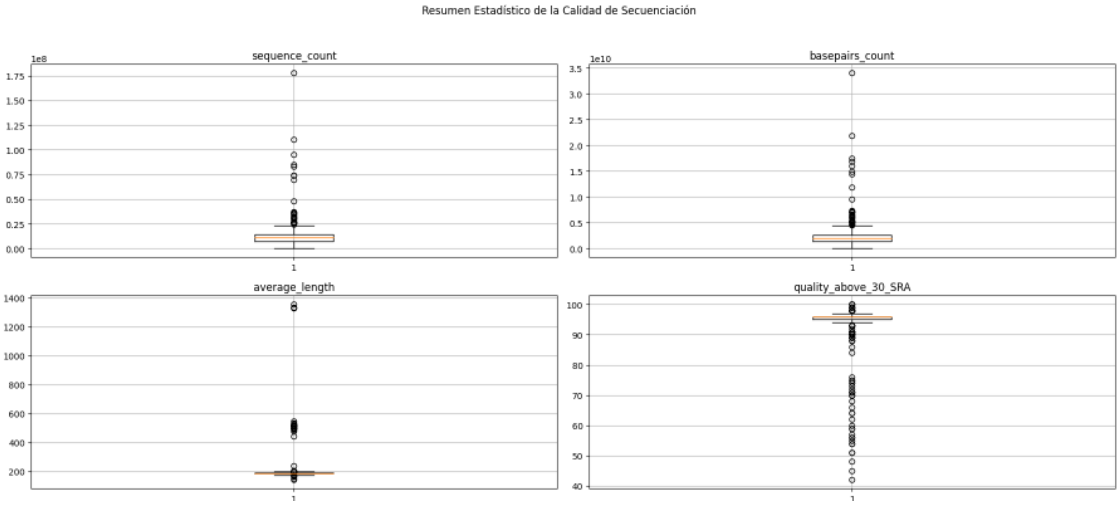
Distribución del centro de investigación. Elaboración propia con librería matplotlib. Python

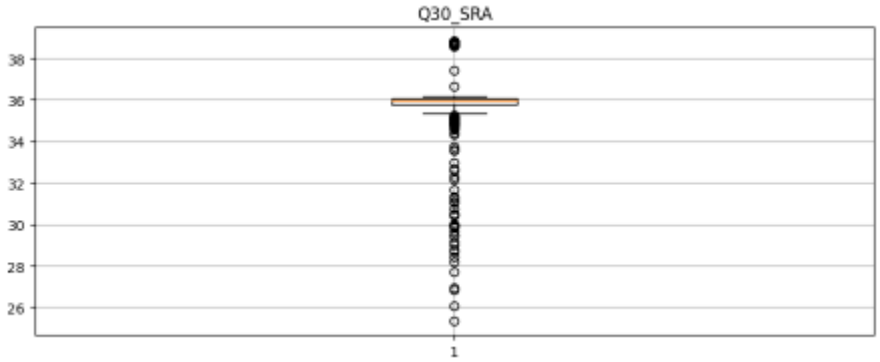


Distribución de la descripción del proyecto para CU MGnify. Elaboración propia con librería matplotlib. Python

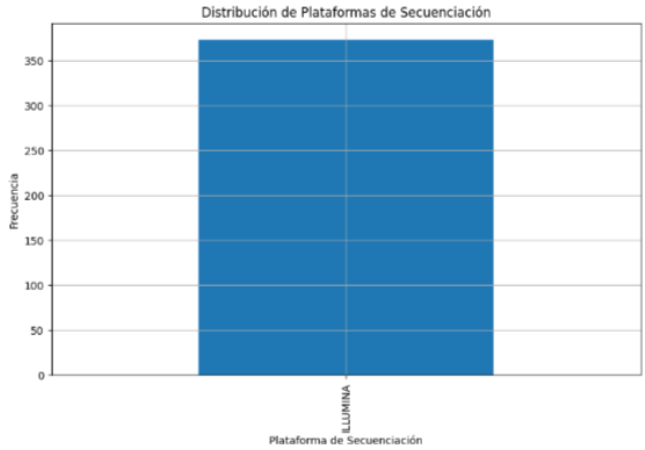


Distribución de plataformas de secuenciación para EC HMgDB. Elaboración propia con librería matplotlib. Python



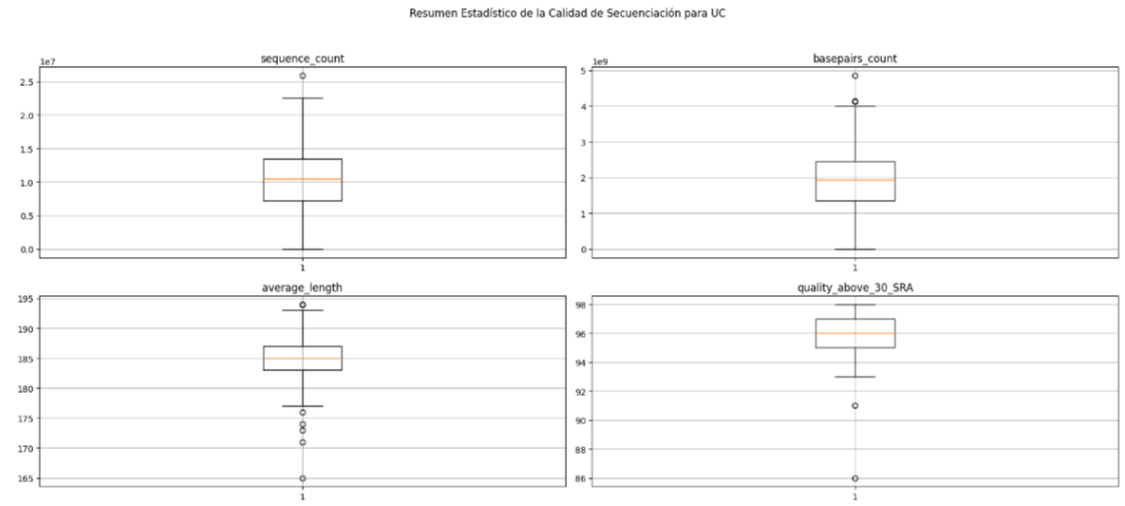


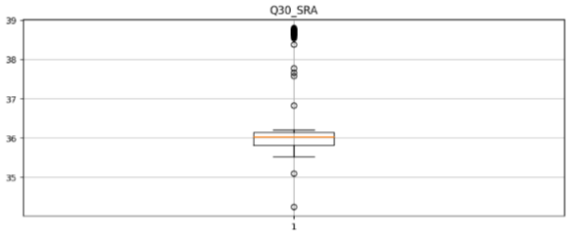

```
seq_platform
ILLUMINA    373
Name: count, dtype: int64
```

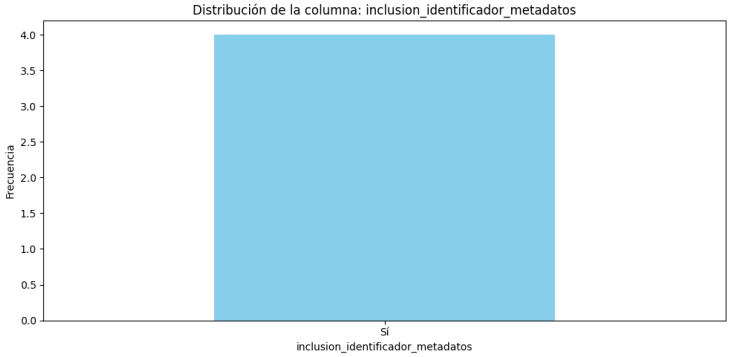


Distribución de plataformas de secuenciación para CU HMgDB. Elaboración propia con librería matplotlib. Python

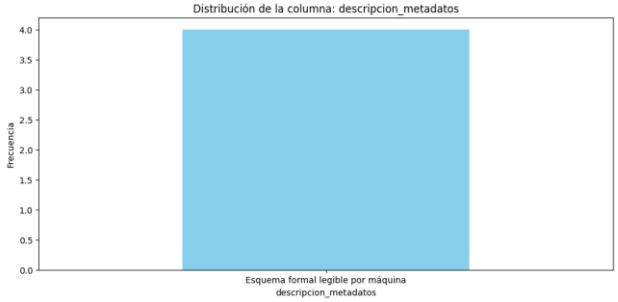
	quality_above_30_SRA	Q30_SRA	creation_date	update_date	source_database
0	95	35.80	2017-08-15	2017-08-15	SRA
1	97	36.14	2017-08-15	2017-08-15	SRA
2	96	36.05	2017-08-15	2017-08-15	SRA
3	94	35.68	2017-08-15	2017-08-15	SRA
4	96	36.10	2017-08-15	2017-08-15	SRA



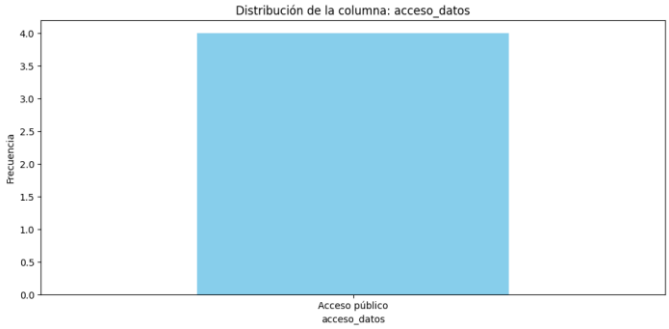




Descripción de metadatos. Elaboración propia con librería matplotlib. Python

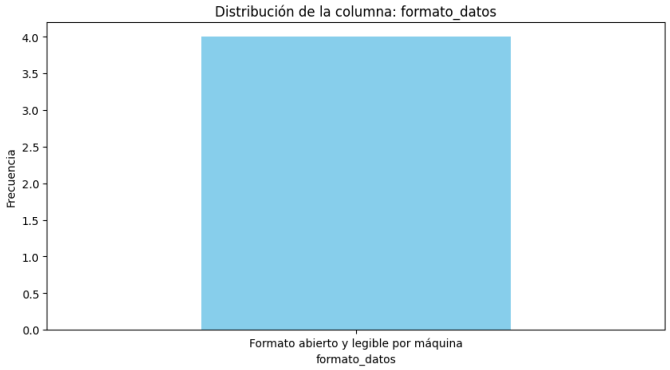


Descripción de metadatos. Elaboración propia con librería matplotlib. Python

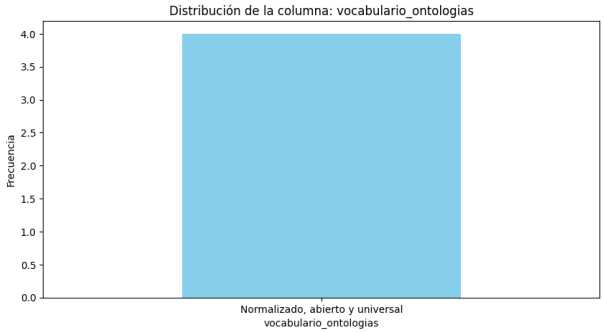


Accesibilidad de los datos. Elaboración propia con librería matplotlib. Python

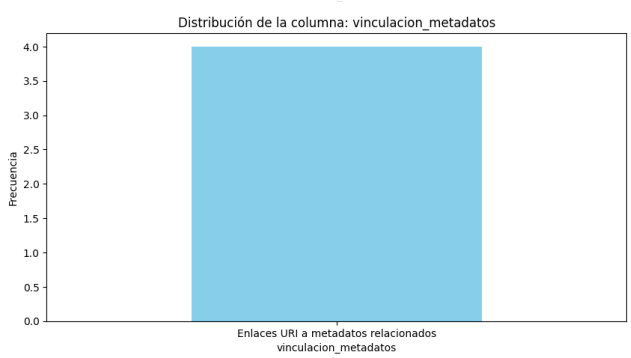
Tipo de repositorio. Elaboración propia con librería matplotlib. Python



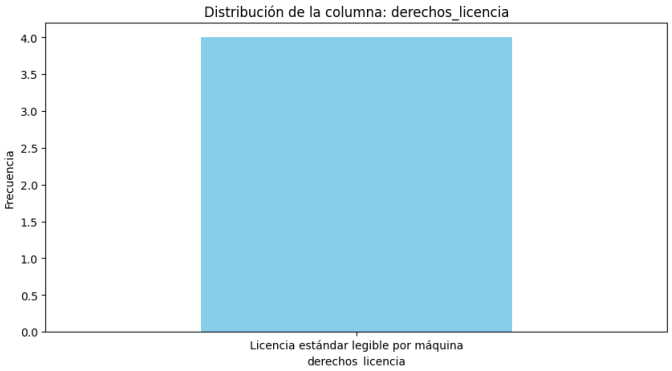
Formato de datos. Elaboración propia con librería matplotlib. Python



Vocabulario ontologías. Elaboración propia con librería matplotlib. Python



Vinculación de metadatos con otros datos y metadatos. Elaboración propia con librería matplotlib. Python



9 Anexo 2: Listado Cotejo IA

Response was added on 24-09-2024 04:57.

Nombre y apellidos	Osman Javier Lòpez Gualdròn
Número de Expediente o Número Empleado	221d0667
Nombre del programa	◯ Máster Universitario en Urgencias, Emergencias y Críticos en Enfermería ⊗ Máster Universitario en Bioinformática
He recibido una explicación satisfactoria sobre el procedimiento del estudio, su finalidad, riesgos y beneficios. He quedado satisfecho/a con la información recibida, la he comprendido y se me han respondido todas mis dudas. Presto mi consentimiento para el procedimiento propuesto	⊗ Yes ◯ No
Marco de Investigación	
Generación de Ideas: Utilizar inteligencia artificial para proponer y desarrollar nuevas ideas o enfoques de investigación	◯ Nivel de uso Bajo (1) No he utilizado herramientas de IA para la generación de ideas. No aplica a este trabajo ⊗ Nivel de uso Medio (2) He aplicado herramientas de IA para generar ideas relacionadas con mi trabajo de investigación, aunque han tenido que ser adaptadas. ◯ Nivel de uso Alto (3) He utilizado herramientas de IA para generar ideas originales y pertinentes y las he aplicado directamente a la realización de mi trabajo.
Generación de Ideas: Herramientas IA usadas	ChatGPT-4o
Generación de Ideas: Utilidad de herramienta (de 0 a 5)	◯ 1 ⊗ 2 ◯ 3 ◯ 4 ◯ 5
Generación de Ideas: Prompt utilizado	Identifica los principales desafíos en la adherencia a los principios FAIR en bases de datos metagenómicas para el estudio de enfermedades inflamatorias intestinales.

Formulación de Hipótesis : Emplear la inteligencia artificial para formular hipótesis claras y fundamentadas que guíen la investigación	○ Nivel de uso Bajo (1) No he utilizado herramientas de IA para la formulación de hipótesis. No aplica a este trabajo ⊗ Nivel de uso Medio (2) He usado herramientas de IA para la identificación de relaciones complejas, pero han sido adaptadas y reformuladas. ○ Nivel de uso Alto (3) He formulado hipótesis de manera clara y precisa utilizando herramientas de IA para identificar relaciones complejas sin hacer ninguna adaptación.
Formulación de Hipótesis: Herramientas IA usadas	ChatGPT-4o
Formulación de Hipótesis : Utilidad de herramienta (de 0 a 5)	○ 1 ⊗ 2 ○ 3 ○ 4 ○ 5
Formulación de Hipótesis : Prompt utilizado	Identifica los principales desafíos en la adherencia de bases de datos metagenómicas a los principios FAIR y describe cómo estas dificultades podrían limitar su uso en modelos de aprendizaje automático.
Diseño de Investigación: Utilizar la inteligencia artificial para diseñar metodologías de investigación sólidas y efectivas	⊗ Nivel de uso Bajo (1) No he utilizado herramientas de IA para el diseño de investigación. No aplica a este trabajo ○ Nivel de uso Medio (2) He usado herramientas de IA para plasmar el diseño de investigación, aunque la información obtenida ha sido reformulada para adaptarse al trabajo. ○ Nivel de uso Alto (3) He diseñado un estudio sólido y bien estructurado, utilizando herramientas de IA.
Diseño de Investigación: Herramientas IA usadas	No aplica
Diseño de Investigación: Utilidad de herramienta (de 0 a 5)	○ 1 ○ 2 ○ 3 ○ 4 ○ 5
Diseño de Investigación: Prompt utilizado	No aplica
Herramientas para Metaanálisis y Revisiones Sistemáticas: Utilizar herramientas de inteligencia artificial para llevar a cabo de manera eficiente y exhaustiva metaanálisis y revisiones sistemáticas en el ámbito de la investigación	○ Nivel de uso Bajo (1) No he utilizado herramientas de IA metaanálisis y revisiones sistemáticas en este trabajo. No aplica a este trabajo ⊗ Nivel de uso Medio (2) He realizado metaanálisis y revisiones sistemáticas que han sido medianamente útiles en la realización de este trabajo. ○ Nivel de uso Alto (3) He realizado metaanálisis y revisiones sistemáticas utilizando herramientas de IA, sin revisar ni contrastar la información.
Herramientas para Metaanálisis y Revisiones Sistemáticas: Herramientas IA usadas	COVIDENCE Y ChatGPT-4o
Herramientas para Metaanálisis y Revisiones Sistemáticas: Utilidad de herramienta (de 0 a 5)	○ 1 ○ 2 ○ 3 ⊗ 4 ○ 5

Herramientas para Metaanálisis y Revisiones Sistemáticas: Prompt utilizado	"Identifica los pasos clave y criterios de inclusión/exclusión necesarios para realizar una revisión sistemática sobre bases de datos metagenómicas utilizando herramientas de automatización como Covidence."
Búsqueda y Sugerencia de Bibliografía: Utilizar la inteligencia artificial para realizar búsquedas bibliográficas avanzadas y encontrar de manera automática sugerencias relevantes de bibliografía para apoyar la investigación	○ Nivel de uso Bajo (1) No he utilizado herramientas de IA para la búsqueda y sugerencia de bibliografía . No aplica a este trabajo ⊗ Nivel de uso Medio (2) He realizado búsquedas bibliográficas adecuadas, pero el uso de herramientas de IA, pero no fueron demasiado útiles o se usaron solo para hacer resúmenes. ○ Nivel de uso Alto (3) He utilizado herramientas de IA para realizar búsquedas exhaustivas y sugerir bibliografía relevante y actualizada, sin revisar ni contrasatar la información
Búsqueda y Sugerencia de Bibliografía: Herramientas IA usadas	ChatGPT-4o
Búsqueda y Sugerencia de Bibliografía: Utilidad de herramienta (de 0 a 5)	○ 1 ⊗ 2 ○ 3 ○ 4 ○ 5
Búsqueda y Sugerencia de Bibliografía: Prompt utilizado	"Identifica términos MeSH y combinaciones de palabras clave para realizar búsquedas avanzadas en PubMed sobre la adherencia a los principios FAIR en bases de datos metagenómicas y su aplicación en modelos de aprendizaje automático."
Localización de Datos y Estadísticas: Utilizar la inteligencia artificial en la localización, recopilación y análisis de datos relevantes y estadísticas necesarias para la investigación	⊗ Nivel de uso Bajo (1) No se han utilizado herramientas de IA para la localización de datos y estadísticas. No aplica a este trabajo ○ Nivel de uso Medio (2) He localizado datos relevantes y aplicado análisis estadísticos aplicando herramientas de IA, pero han sido medianamente útiles y han requerido búsqueda manual. ○ Nivel de uso Alto (3) He utilizado herramientas de IA para localizar datos relevantes y aplicar análisis estadísticos avanzados, sin revisar ni analizar la información.
Localización de Datos y Estadísticas: Herramientas IA usadas	No aplica
Localización de Datos y Estadísticas: Utilidad de herramienta (de 0 a 5)	○ 1 ○ 2 ○ 3 ○ 4 ○ 5
Localización de Datos y Estadísticas: Prompt utilizado	No aplica

Análisis de datos : Emplear la inteligencia artificial para realizar análisis avanzados de datos y generar resultados significativos y precisos	⊗ Nivel de uso Bajo (1) No he utilizado herramientas de IA para realizar Análisis de datos. No aplica a este trabajo ○ Nivel de uso Medio (2) He utilizado herramientas de IA para generar ideas sobre el análisis de datos, verificando la información proporcionada por la herramienta mediante la comparación de los resultados con análisis manuales o con fuentes externas relevantes ○ Nivel de uso Alto (3) He utilizado herramientas de IA para realizar el análisis de datos sin verificar que los resultados proporcionados por la herramienta son correctos.
Análisis de datos : Herramientas IA usadas	No aplica
Análisis de datos : Utilidad de herramienta (de 0 a 5)	○ 1 ○ 2 ○ 3 ○ 4 ○ 5
Análisis de datos : Prompt utilizado	No aplica
Resumen de Información: Utilizar la inteligencia artificial en la síntesis y resumen de información relevante obtenida durante la investigación, facilitando la comunicación y comprensión de los hallazgos	○ Nivel de uso Bajo (1) No se han utilizado herramientas de IA para resumir información. No aplica a este trabajo ⊗ Nivel de uso Medio (2) He usado herramientas de IA para sintetizar la información, pero la salida de los resúmenes ha tenido que ser reescrita. ○ Nivel de uso Alto (3) He sintetizado de manera efectiva la información obtenida utilizando herramientas de IA para identificar patrones y tendencias, sin revisar ni analizar la información
Resumen de Información: Herramientas IA usadas	ChatGPT-4o
Resumen de Información: Utilidad de herramienta (de 0 a 5)	○ 1 ⊗ 2 ○ 3 ○ 4 ○ 5
Resumen de Información: Prompt utilizado	"Resumir los principales hallazgos y conclusiones de la investigación sobre la adherencia a los principios FAIR en bases de datos metagenómicas, destacando puntos clave para facilitar la comunicación y comprensión de los resultados."
Redacción Académica , Edición y Revisión	
Traducción del lenguaje: Utilizar la inteligencia artificial para traducir texto de un idioma a otro de manera precisa y coherente en el contexto académico, ya sean para la traducción de artículos o documentos científicos, o para la traducción de determinadas partes del trabajo.	⊗ Nivel de uso Bajo (1) No se han utilizado herramientas de IA para traducir el lenguaje. No aplica a este trabajo ○ Nivel de uso Medio (2) He utilizado herramientas de IA para traducir directamente a otro idioma fragmentos de otras fuentes de información y elaborar una nueva redacción ○ Nivel de uso Alto (3) He utilizado herramientas de IA para traducir directamente a otro idioma fragmentos de otras fuentes de información
Traducción del lenguaje: Herramientas IA usadas	No aplica

Traducción del lenguaje: Utilidad de herramienta (de 0 a 5)	○ 1 ○ 2 ○ 3 ○ 4 ○ 5
Traducción del lenguaje: Prompt utilizado	No aplica
Interpretación de la información: Utilizar la inteligencia artificial para analizar y entender información compleja y detallada encontrada en documentación científica, permitiendo una comprensión más profunda de los datos y resultados	○ Nivel de uso Bajo (1) No se han utilizado herramientas de IA para interpretar la información ⊗ Nivel de uso Medio (2)He utilizado herramientas de IA para interpretar los resultados de otras fuentes de información y adaptarlos a mi trabajo . ○ Nivel de uso Alto (3) He utilizado herramientas de IA para interpretar la información de otras fuentes de información con mi trabajo, sin contrastar la información
Interpretación de la información: Herramientas IA usadas	ChatGPT-4o
Interpretación de la información: Utilidad de herramienta (de 0 a 5)	○ 1 ○ 2 ○ 3 ⊗ 4 ○ 5
Interpretación de la información: Prompt utilizado	"Analiza y explica los resultados de los estudios sobre la adherencia a los principios FAIR en bases de datos metagenómicas, destacando cómo estos resultados pueden influir en la investigación y aplicación en modelos de aprendizaje automático."
Escritura o convertidor a texto académico (enriquecer vocabulario, verificar gramática, corrección ortográfica, etc): Emplear la inteligencia artificial para redactar o convertir texto en un formato académico, enriqueciendo el vocabulario, corrigiendo la gramática y la ortografía de manera automatizada y precisa	○ Nivel de uso Bajo (1) No se han utilizado herramientas de IA para escribir el trabajo. No aplica a este trabajo ⊗ Nivel de uso Medio (2) He utilizado herramientas de IA para redactar algunos fragmentos del trabajo para darle un enfoque científico, haciendo una revisión personal para asegurar la claridad y coherencia ○ Nivel de uso Alto (3) He utilizado herramientas de IA para mejorar la calidad de la redacción, corrección de errores gramaticales, sugerir mejoras de estilo y mantener la consistencia y el tono formal y académico, sin hacer revisión final
Escritura o convertidor a texto académico : Herramientas IA usadas	ChatGPT-4o
Escritura o convertidor a texto académico : Utilidad de herramienta (de 0 a 5)	○ 1 ○ 2 ○ 3 ⊗ 4 ○ 5
Escritura o convertidor a texto académico : Prompt utilizado	"Revisa y mejora el texto para cumplir con los estándares académicos, enriqueciendo el vocabulario, verificando la gramática y corrigiendo la ortografía, manteniendo un estilo formal y coherente."

Parafraseo :Utilizar la inteligencia artificial para reformular y parafrasear texto de manera coherente, preservando el significado original, pero utilizando diferentes palabras o estructuras gramaticales	○ Nivel de uso Bajo (1) No se han utilizado herramientas de IA para parafrasear otros textos. No aplica a este trabajo ⊗ Nivel de uso Medio (2) He utilizado herramientas de IA para parafrasear pequeñas ideas de la fuente original, pero haciendo aportaciones personales en la redacción final ○ Nivel de uso Alto (3) He parafraseado fragmentos completos que son copia literal de la fuente utilizando herramientas de IA
Parafraseo : Herramientas IA usadas	ChatGPT-4o
Parafraseo : Utilidad de herramienta (de 0 a 5)	○ 1 ○ 2 ⊗ 3 ○ 4 ○ 5
Parafraseo : Prompt utilizado	"Reformula el siguiente texto manteniendo el significado original, utilizando diferentes palabras y estructuras gramaticales para aumentar la fluidez. Asegúrate de que el texto resultante tenga una alta explosividad y perplejidad, imitando la variabilidad y estilo de la escritura humana."
Gestión de referencias bibliográficas y verificación de citas : Utilizar la inteligencia artificial para gestionar y organizar eficientemente las referencias bibliográficas en el trabajo académico, así como verificar la precisión de las citas según los estándares de formato establecidos	⊗ Nivel de uso Bajo (1) NNo se han utilizado herramientas de IA para gestionar referencias bibliográficas ni citas. No aplica a este trabajo ○ Nivel de uso Medio (2) He utilizado herramientas de IA para hacer una primera aproximación en la búsqueda bibliográfica contrastando la existencia y la relevancia de los artículos en bases de datos científicas y revisando las citas ○ Nivel de uso Alto (3) He utilizado herramientas de IA para hacer búsquedas bibliográficas y gestionar las citas en el trabajo sin contrastar la información
Gestión de referencias bibliográficas y verificación de citas : Herramientas IA usadas	No aplica
Gestión de referencias bibliográficas y verificación de citas : Utilidad de herramienta (de 0 a 5)	○ 1 ○ 2 ○ 3 ○ 4 ○ 5
Gestión de referencias bibliográficas y verificación de citas: Prompt utilizado	No aplica
Generación del título o palabras clave : Utilizar la inteligencia artificial para generar títulos relevantes y palabras clave que representen con precisión el contenido del trabajo académico	⊗ Nivel de uso Bajo (1) No se han utilizado herramientas de IA para elebaorar el título ni defnir las plabaras clave. No aplica a este trabajo ○ Nivel de uso Medio (2) He utilizado herramientas de IA para la generación del título, definiendo el título final con mi aportación personal ○ Nivel de uso Alto (3) He utilizado herramientas de IA para la generación del título y la selección de palabras clave del trabajo, sin hacer una revisión y análisis de la información
Generación del título o palabras clave : Herramientas IA usadas	No aplica

Generación del título o palabras clave : Utilidad de herramienta (de 0 a 5)	○ 1 ○ 2 ○ 3 ○ 4 ○ 5
Generación del título o palabras clave: Prompt utilizado	No aplica
Verificación de consistencia de las conclusiones del trabajo : Emplear la inteligencia artificial para redactar conclusiones claras y coherentes en el trabajo académico, así como para verificar la consistencia y validez de las conclusiones en relación a los objetivos propuestos y los resultados obtenidos	○ Nivel de uso Bajo (1) No se han utilizado herramientas de IA para verificar la consistencia de las conclusiones del trabajo. No aplica a este trabajo. ⊗ Nivel de uso Medio (2) He utilizado herramientas de IA para verificar la consistencia del trabajo, redefiniéndolas con aportaciones personales ○ Nivel de uso Alto (3) He utilizado herramientas de IA para verificar la consistencia de las conclusiones del trabajo, sin hacer una revisión y análisis de la información
Verificación de consistencia de las conclusiones del trabajo : Herramientas IA usadas	ChatGPT-4o
Verificación de consistencia de las conclusiones del trabajo : Utilidad de herramienta (de 0 a 5)	○ 1 ⊗ 2 ○ 3 ○ 4 ○ 5
Verificación de consistencia de las conclusiones del trabajo: Prompt utilizado	"Revisa y asegura que las conclusiones del trabajo sean claras, coherentes y alineadas con los objetivos propuestos y los resultados obtenidos. Verifica la validez de las conclusiones, destacando cualquier inconsistencia y proponiendo ajustes si es necesario."
Análisis y visualización de datos	
Análisis de datos: Emplear la inteligencia artificial para examinar, procesar y comprender los datos durante el trabajo, identificando tendencias o relaciones y responder preguntas o tomar decisiones	○ Nivel de uso Bajo (1) No he utilizado herramientas de IA para realizar Análisis de datos. No aplica a este trabajo ⊗ Nivel de uso Medio (2) He utilizado herramientas de IA para generar ideas sobre el análisis de datos, verificando la información proporcionada por la herramienta mediante la comparación de los resultados con análisis manuales o con fuentes externas relevantes ○ Nivel de uso Alto (3) He utilizado herramientas de IA para realizar el análisis de datos sin verificar que los resultados proporcionados por la herramienta son correctos
Análisis de datos: Herramientas IA usadas	ChatGPT-4o
Análisis de datos: Utilidad de herramienta (de 0 a 5)	○ 1 ○ 2 ⊗ 3 ○ 4 ○ 5
Análisis de datos: Prompt utilizado	"Examina y procesa los datos para identificar tendencias y relaciones clave, resumiendo los hallazgos de manera comprensible. Proporciona un análisis detallado que responda a las preguntas de investigación y sugiere decisiones basadas en los resultados observados."

Análisis estadístico: Utilizar la inteligencia artificial para realizar análisis estadísticos avanzados de los datos, incluyendo medidas descriptivas, pruebas de hipótesis y análisis multivariados o predictivos.	○ Nivel de uso Bajo (1) No he utilizado herramientas de IA para realizar Análisis estadístico. No aplica a este trabajo ⊗ Nivel de uso Medio (2) He localizado datos relevantes y aplicado análisis estadísticos utilizando herramientas de IA. Para garantizar la precisión de los resultados, he verificado los resultados obtenidos comparándolos con análisis estadísticos manuales y corroborando la coherencia con conocimientos previos. ○ Nivel de uso Alto (3) He utilizado herramientas de IA para realizar los análisis estadísticos avanzados, sin revisar ni verificar la información
Análisis estadístico : Herramientas IA usadas	ChatGPT-4o
Análisis estadístico: Utilidad de herramienta (de 0 a 5)	○ 1 ○ 2 ⊗ 3 ○ 4 ○ 5
Análisis estadístico: Prompt utilizado	"Guíame paso a paso en la realización de un análisis estadístico avanzado utilizando Python. Proporciona la estructura y el código necesario para realizar un análisis descriptivo, incluyendo pruebas de hipótesis y análisis multivariado, utilizando las librerías pandas y matplotlib. Asegúrate de comentar detalladamente cada parte del código para que cualquier persona que lo lea pueda entender su funcionamiento y propósito."
Modelado de datos: Utilizar la inteligencia artificial para construir modelos matemáticos o computacionales que representen y expliquen los datos, y permitan crear una representación visual o esquema de los mismos.	⊗ Nivel de uso Bajo (1) No he utilizado herramientas de IA durante el proceso de Modelado de datos. No aplica a este trabajo ○ Nivel de uso Medio (2) He utilizado herramientas de IA para realizar el modelado de datos. Durante este proceso, las herramientas de IA me han ayudado a solucionar problemas técnicos y conceptuales que surgieron, proporcionando soluciones alternativas y sugerencias para mejorar la calidad del modelo. ○ Nivel de uso Alto (3) He utilizado herramientas de IA para realizar el modelado de datos, sin revisar ni verificar la información
Modelado de datos : Herramientas IA usadas	No aplica
Modelado de datos: Utilidad de herramienta (de 0 a 5)	○ 1 ○ 2 ○ 3 ○ 4 ○ 5
Modelado de datos: Prompt utilizado	No aplica

Detección análisis patrones y grandes conjuntos de datos: Utilizar inteligencia artificial para identificar relaciones, tendencias y correlaciones significativas mediante técnicas de Big Data como algoritmos de aprendizaje automático, minería de datos o Machine Learning para extraer conocimientos significativos.	⊗ Nivel de uso Bajo (1) No he utilizado herramientas de IA para realizar la detección de patrones y el análisis de grandes conjuntos de datos No aplica a este trabajo ○ Nivel de uso Medio (2) He utilizado herramientas de IA para generar ideas y corregir problemas durante la detección de patrones y análisis de grandes conjuntos de datos, identificado patrones inesperados que surgieron, proporcionando soluciones alternativas y sugerencias para mejorar la calidad del modelo. ○ Nivel de uso Alto (3) He utilizado herramientas de IA para realizar la detección de patrones y el análisis de grandes conjuntos de datos, sin revisar ni verificar la información
Detección análisis patrones y grandes conjuntos de datos: Herramientas IA usadas	No aplica
Detección análisis patrones y grandes conjuntos de datos: Utilidad de herramienta (de 0 a 5)	○ 1 ○ 2 ○ 3 ○ 4 ○ 5
Detección análisis patrones y grandes conjuntos de datos: Prompt utilizado	No aplica
Simulaciones: Emplear la inteligencia artificial para realizar simulaciones o experimentos a pequeña escala que representen fenómenos o procesos relevantes y permita comprender el comportamiento de una variable a gran escala	⊗ Nivel de uso Bajo (1) No he utilizado herramientas de IA para realizar Simulaciones. No aplica a este trabajo ○ Nivel de uso Medio (2) He utilizado herramientas de IA para obtener ideas y mejorar las simulaciones realizadas. Estas herramientas han proporcionado sugerencias sobre parámetros de entrada, métodos de análisis y alternativas para abordar posibles limitaciones en los modelos simulados ○ Nivel de uso Alto (3) He utilizado herramientas de IA para realizar Simulaciones, sin revisar ni verificar la información
Simulaciones : Herramientas IA usadas	No aplica
Simulaciones: Utilidad de herramienta (de 0 a 5)	○ 1 ○ 2 ○ 3 ○ 4 ○ 5
Simulaciones: Prompt utilizado	No aplica
Modelado predictivo: Utilizar la inteligencia artificial para desarrollar modelos predictivos estadísticos o de aprendizaje automático que puedan anticipar resultados o comportamientos de eventos futuros basados en datos históricos o en tiempo real en el contexto de la investigación	⊗ Nivel de uso Bajo (1) No he utilizado herramientas de IA para realizar modelos predictivos. No aplica a este trabajo ○ Nivel de uso Medio (2) He utilizado herramientas de IA para obtener ideas y corregir errores durante la creación de modelos predictivos en los datos de entrenamiento y sugeridos ajustes en los algoritmos de predicción para mejorar la precisión del modelo. ○ Nivel de uso Alto (3) He utilizado herramientas de IA para realizar modelos predictivos, proporcionando datos, sin revisar ni verificar la información

Modelado predictivo : Herramientas IA usadas	No aplica
Modelado predictivo : Utilidad de herramienta (de 0 a 5)	○ 1 ○ 2 ○ 3 ○ 4 ○ 5
Modelado predictivo : Prompt utilizado	No aplica
Creación de gráficos, tablas, esquemas, etc: Utilizar la inteligencia artificial para generar representaciones gráficas de la información y los datos, facilitando la comprensión e interpretación de la información y los resultados obtenidos.	○ Nivel de uso Bajo (1) No he utilizado herramientas de IA para la creación de gráficos, tablas, esquemas, etc. No aplica a este trabajo ⊗ Nivel de uso Medio (2) He utilizado herramientas de IA para obtener ideas y sugerencias de diseño para la creación de los gráficos mostrados ○ Nivel de uso Alto (3) He utilizado herramientas de IA para la creación de gráficos sin hacer ninguna modificación
Creación de gráficos, tablas, esquemas, etc. : Herramientas IA usadas	ChatGPT-4o
Creación de gráficos, tablas, esquemas, etc: Utilidad de herramienta (de 0 a 5)	○ 1 ○ 2 ⊗ 3 ○ 4 ○ 5
Creación de gráficos, tablas, esquemas, etc: Prompt utilizado	"Proporciona una guía paso a paso para crear representaciones gráficas de los datos utilizando Python, específicamente con las librerías pandas y matplotlib. Incluye instrucciones detalladas y ejemplos de código para generar gráficos, tablas y esquemas, explicando cada parte del código para facilitar la comprensión y la interpretación de los resultados obtenidos."
Generación de imágenes: Emplear la inteligencia artificial para generar imágenes mediante algoritmos avanzados de aprendizaje automático que permiten generar representaciones visuales basadas en datos de entrada	⊗ Nivel de uso Bajo (1) No he utilizado herramientas de IA para generar imágenes. No aplica a este trabajo ○ Nivel de uso Medio (2) He utilizado herramientas de IA para obtener ideas y sugerencias de diseño para la creación de las imágenes. ○ Nivel de uso Alto (3) He utilizado herramientas de IA para Generar imágenes, sin hacer ninguna modificación
Generación de imágenes : Herramientas IA usadas	No aplica
Generación de imágenes : Utilidad de herramienta (de 0 a 5)	○ 1 ○ 2 ○ 3 ○ 4 ○ 5
Generación de imágenes: Prompt utilizado	No aplica