Taller 2: BASES DE DATOS Y BÚSQUEDAS DE INFORMACIÓN BIOLÓGICA

Carlos Andrés Díaz - **código: 202010343**

David León - **código: 201615216**

Cesar Patiño - **código: 201924259**

**"palabraclave"[field] <operador lógico> "palabraclave"[field] …**

**Nota:** Si no está seguro de qué campos elegir para una búsqueda dada, haga click en“*Advanced*” debajo de la barra de búsqueda. Esto lo lleva a una interfaz que le permite construir su búsqueda personalizada. Al momento de buscar, en la barra quedará consignado cómo debería escribirse la misma búsqueda usando campos y operadores booleanos directamente. Puede encontrar más ayuda aquí <https://www.ncbi.nlm.nih.gov/books/NBK3827/#pubmedhelp.Advanced_Search>

**IMPORTANTE:** Especifique como realizo la búsqueda para cada punto. Si es necesario adjunte pantallazos que acompañen la búsqueda.

**Para entregar:**

En el **NCBI**, haga una búsqueda de literatura en <https://pubmed.ncbi.nlm.nih.gov/>

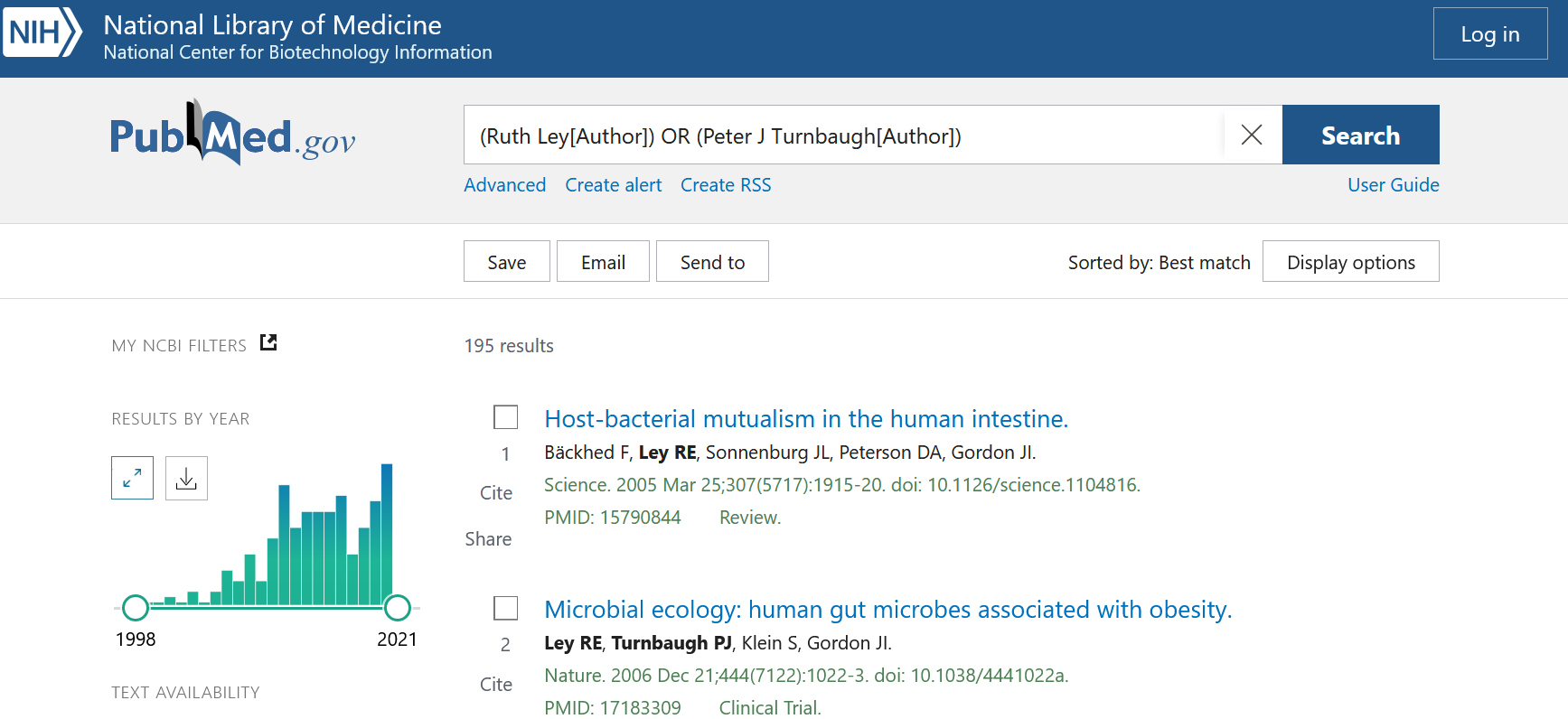
* Busque artículos con Ruth Ley como autor. Cuantos artículos encuentra?

**Respuesta:** Al utilizar la ecuación de búsqueda *“Ley, Ruth [Author] “*, Se encontraron 111 artículos publicados por esta autora



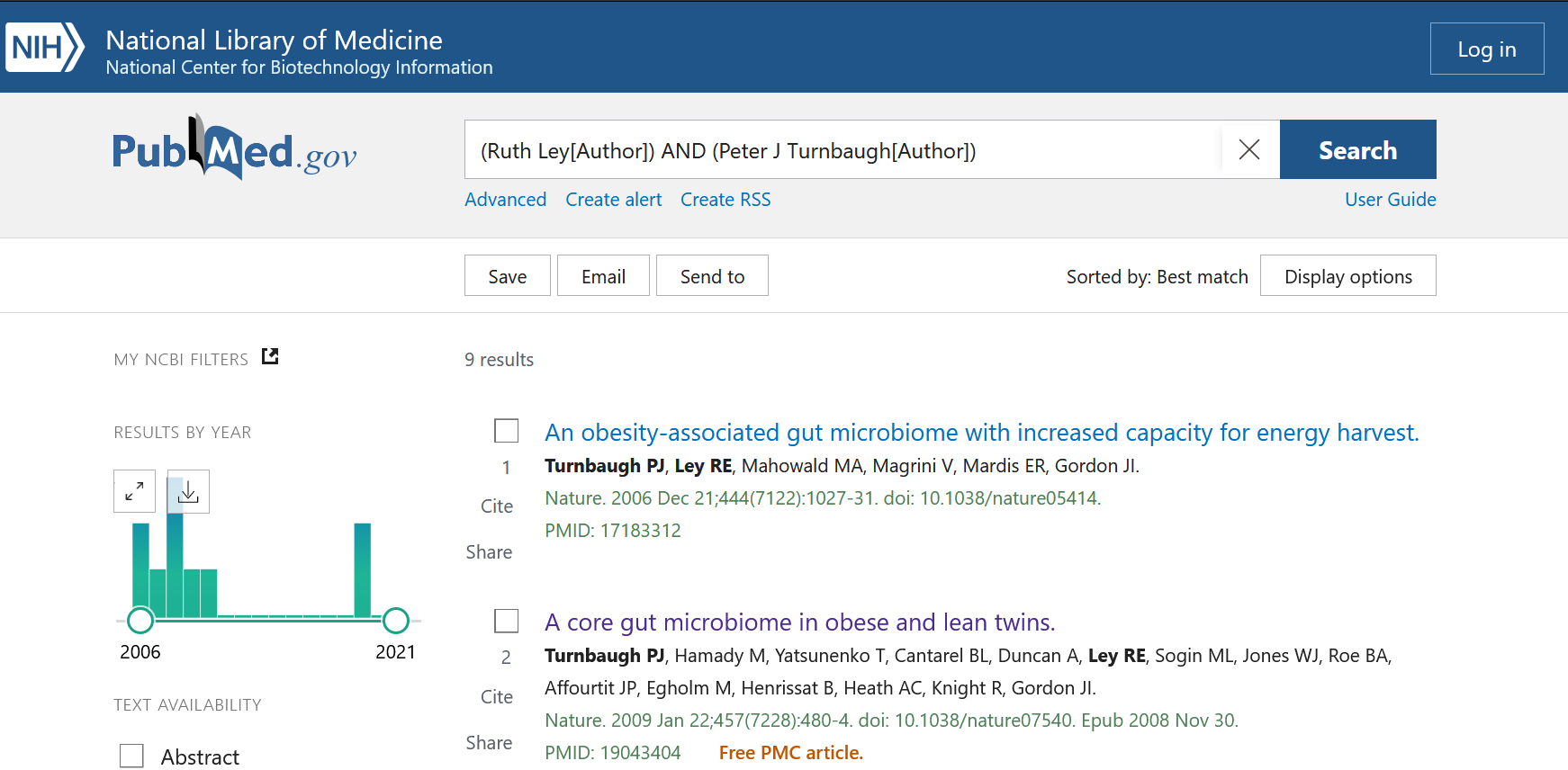
* Ahora busque artículos por Ruth Ley o Peter J Turnbaugh, cuantos obtiene?

**Respuesta:** Al utilizar la ecuación de búsqueda *“(Ruth Ley[Author]) OR (Peter J Turnbaugh[Author]) “*, Se encontraron 195 artículos publicados por ambos autores



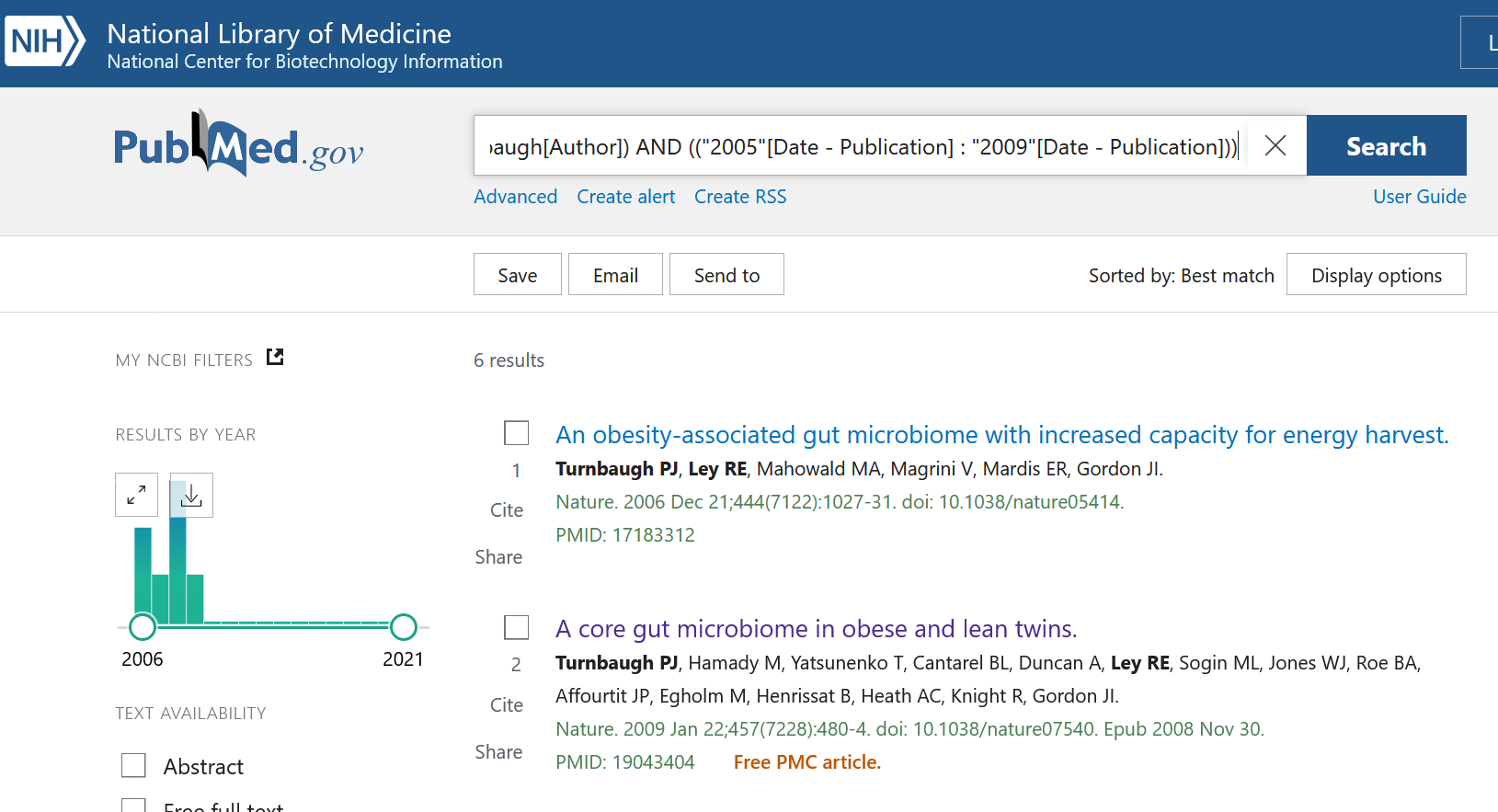
* Ahora cuantos artículos fueron escritos por ambos.

**Respuesta:** Al utilizar la ecuación de búsqueda *“(Ruth Ley[Author]) AND (Peter J Turnbaugh[Author]) “*, Se encontraron 9 artículos publicados por ambos autores



* Y si limitamos solo al tiempo en que estuvieron trabajando juntos (2005-2009). ¿Cuantos se obtienen?

**Respuesta:** Al utilizar la ecuación de búsqueda *“(Ruth Ley[Author]) AND (Peter J Turnbaugh[Author]) AND (("2005"[Date - Publication] : "2009"[Date - Publication])) “*, Se encontraron 6 artículos publicados por ambos autores durante ese periodo de tiempo



Entre los artículos publicados por ellos uno de los más clásicos es el análisis del microbioma de gemelos delgados y obesos publicado en Nature en 2009. Seleccione ese artículo.

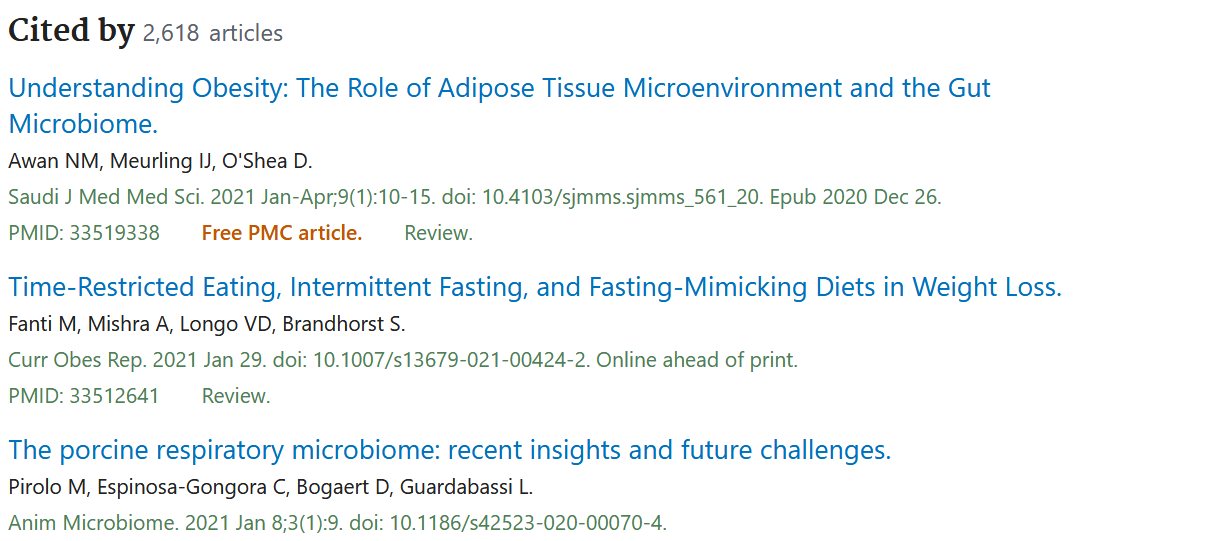
* ¿Cuantos individuos analizaron en ese estudio? Cuantas secuencias de 16S de longitud completa y cuantas parciales? Adicionalmente cuanta información del microbioma se generó?

**Respuesta:** Se analizaron un número total de 154 Individuos, además se encontraron 9,920 secuencias completas y 1,937,461 secuencias parciales de RNA ribosomal 16S. Al realizar la secuenciación se obtuvieron 2.14 Gigabases



* ¿Cuántas citas ha obtenido este artículo a la fecha?

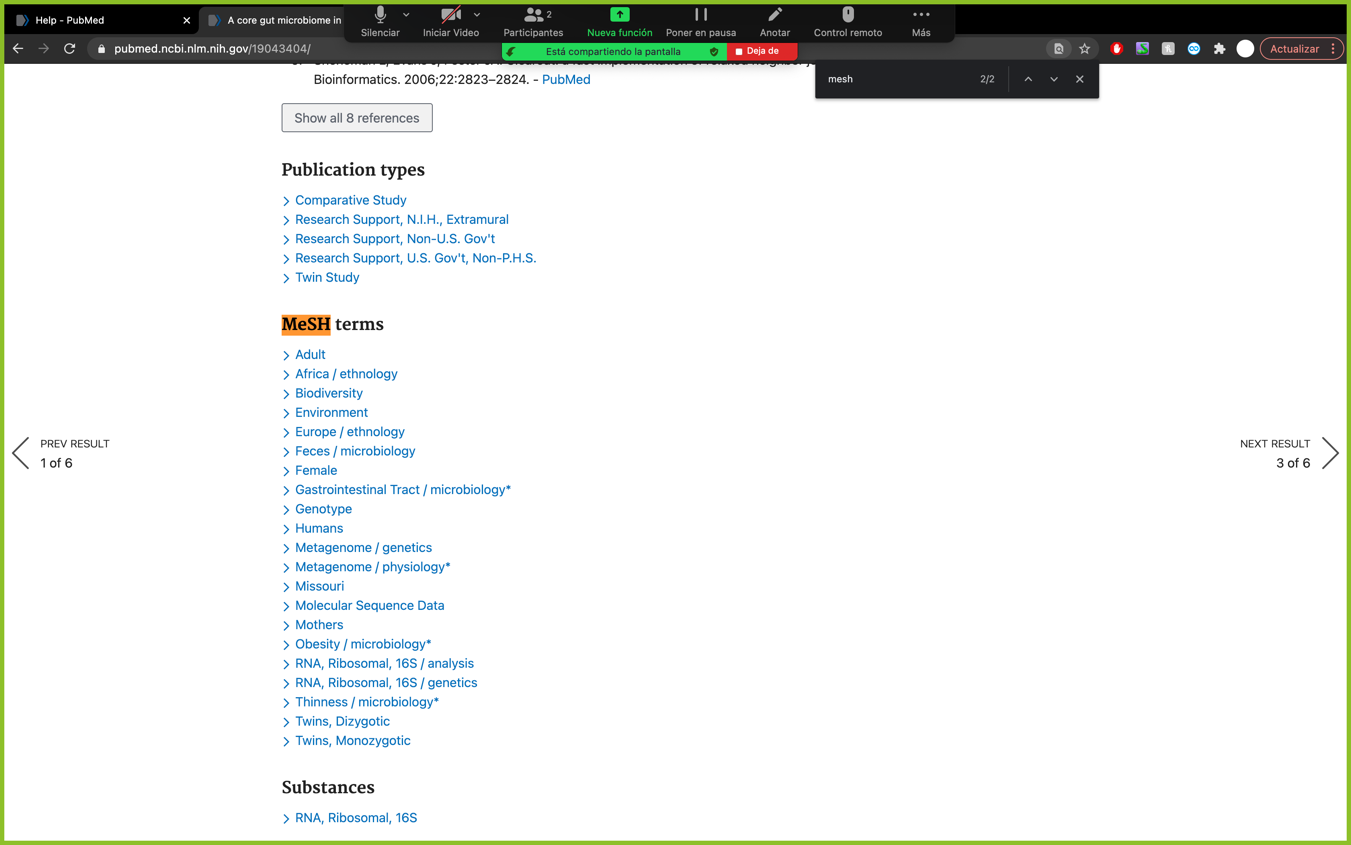
**Respuesta:** Al utilizar la seleccionar el artículo, observamos que se encuentra citado en 2,618 artículos científicos, tal y como lo muestra la siguiente figura:



Entre los términos MeSH usados uno de los que se nombra es Metagenome / genetics.

¿Qué es un término MeSH? Y ¿Cuántos artículos se han publicado con el termino MeSH Metagenome / genetics

**Respuesta:** “El término MeSH hace referencia a los «medical subject headings», también conocidos como «encabezamientos de materia médicos», «títulos de temas médicos» o «descriptores en ciencias de la salud», los cuales son utilizados por la NLM.” (Fernández-Altuna et al., 2016)



* Cuantos artículos se han publicado con el termino MeSH Metagenome / genetics antes de la fecha de publicación de este artículo? Considera que este fue uno de los artículos pioneros en el campo? Por qué?

**Respuesta:** Se encontraron 12 artículos con el término MeSH Metagenome / genetics antes de la fecha de publicación del artículo (22/01/2009).

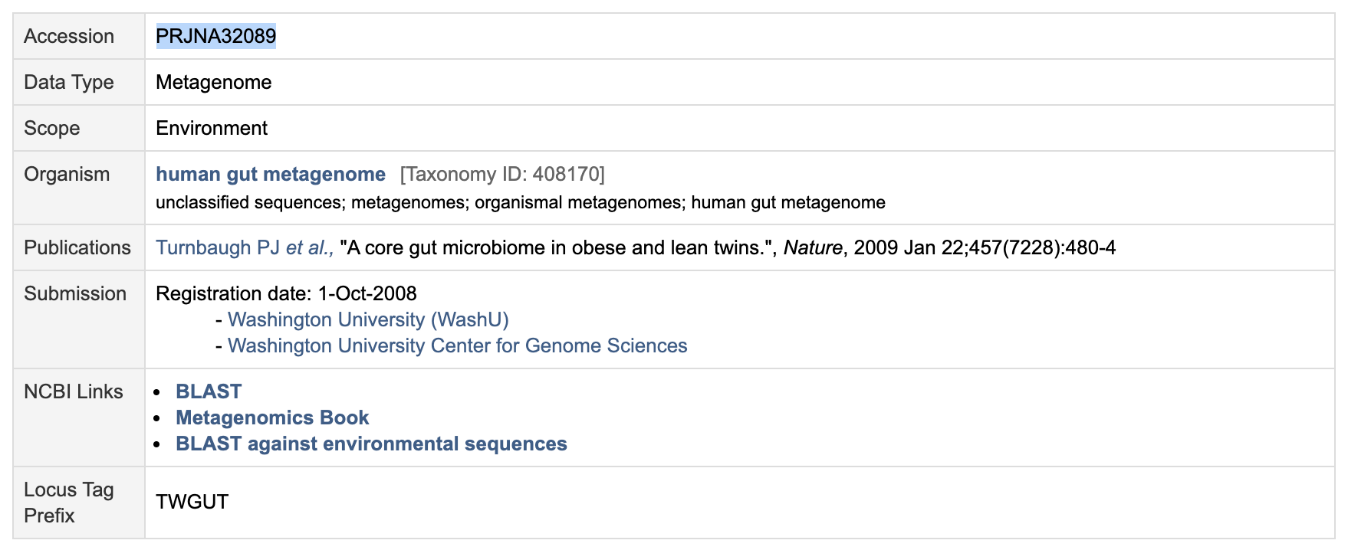
Sí consideramos que fue uno de los artículos pioneros en el campo porque está entre los primeros 15 publicados, además, al revisar los artículos posteriores nos damos cuenta que hay un crecimiento acelerado, por lo que en los últimos años se han añadido más de 1000 artículos con el mismo MeSH



Ahora vamos a buscar los datos.

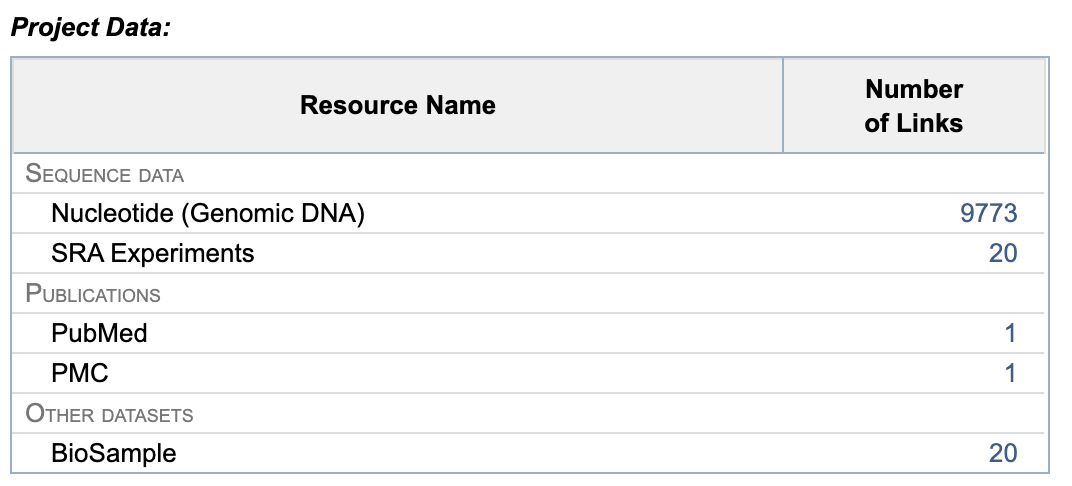
* En Información relacionada, buscar Proyecto Relacionado. ¿Cuál es el número de acceso de este proyecto?

**Respuesta:** El número de acceso de este proyecto es: PRJNA32089 tal y como se muestra en la siguiente figura



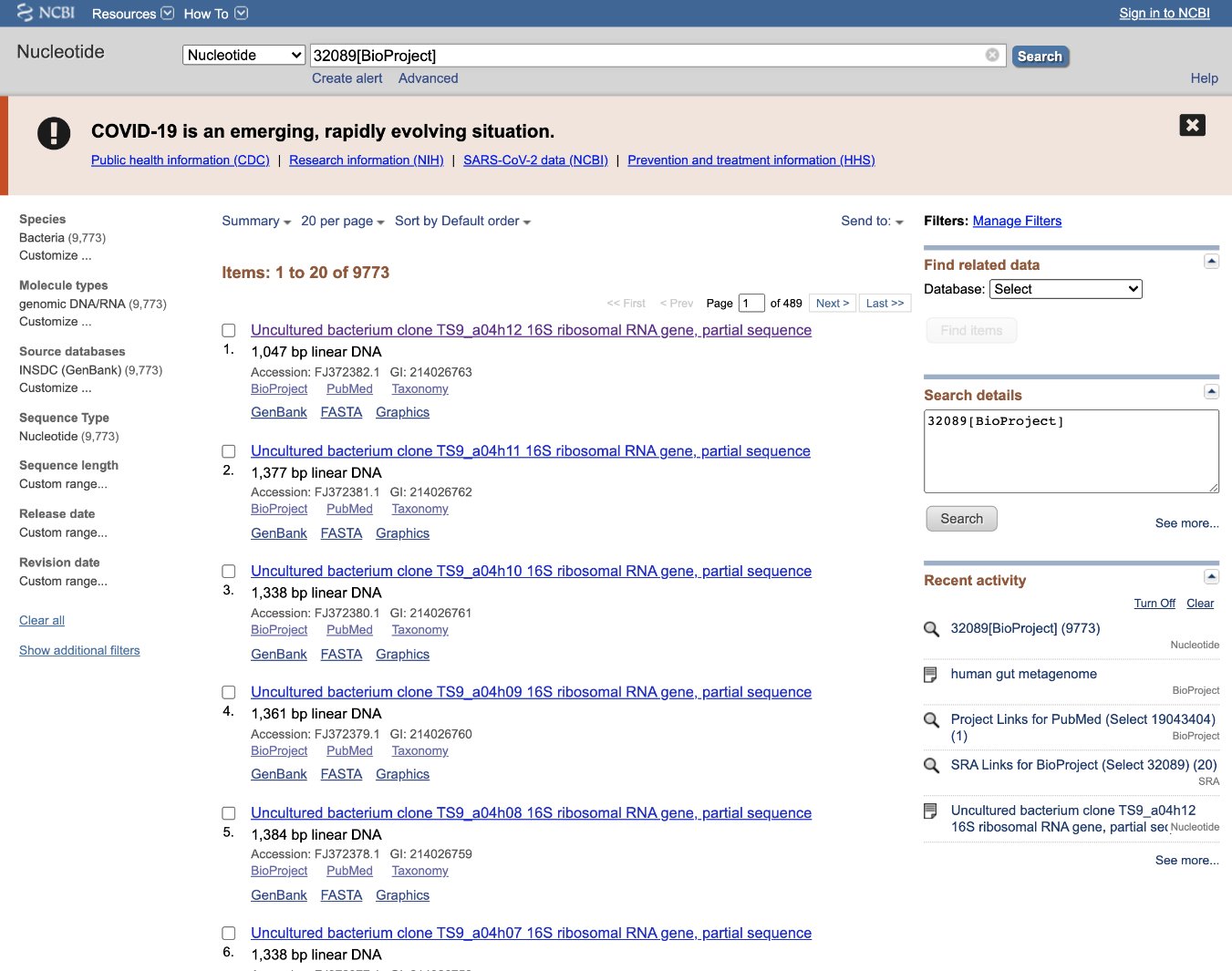
* ¿En cuanto a datos de secuenciación se encuentran tanto en Nucleótidos como en SRA. Cuántas entradas hay en cada uno de estos?

**Respuesta**: Se encuentran 9773 entradas para Nucleótidos y 20 entradas para SRA y 20 entradas para BioSample.



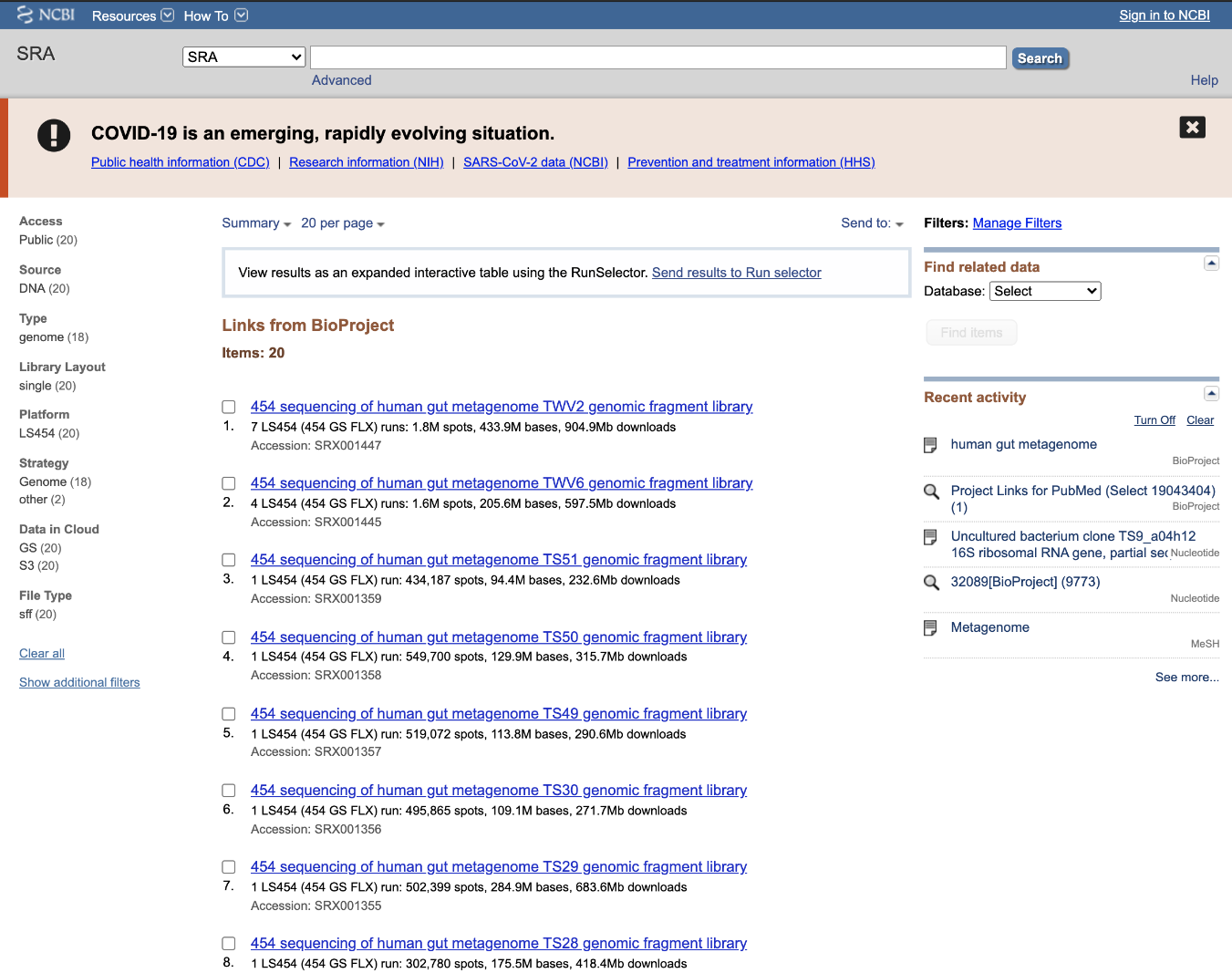
* Al entrar a los datos en Nucleotide, y dada la información sobre secuenciación reportada en el estudio, ¿a qué información considera que corresponde?

**Respuesta:** La información corresponde a las secuencias parciales del gen ribosomal del ARN 16S de cada una de las bacterias, que no se pueden cultivar que conforman el microbioma de los gemelos y las madres.



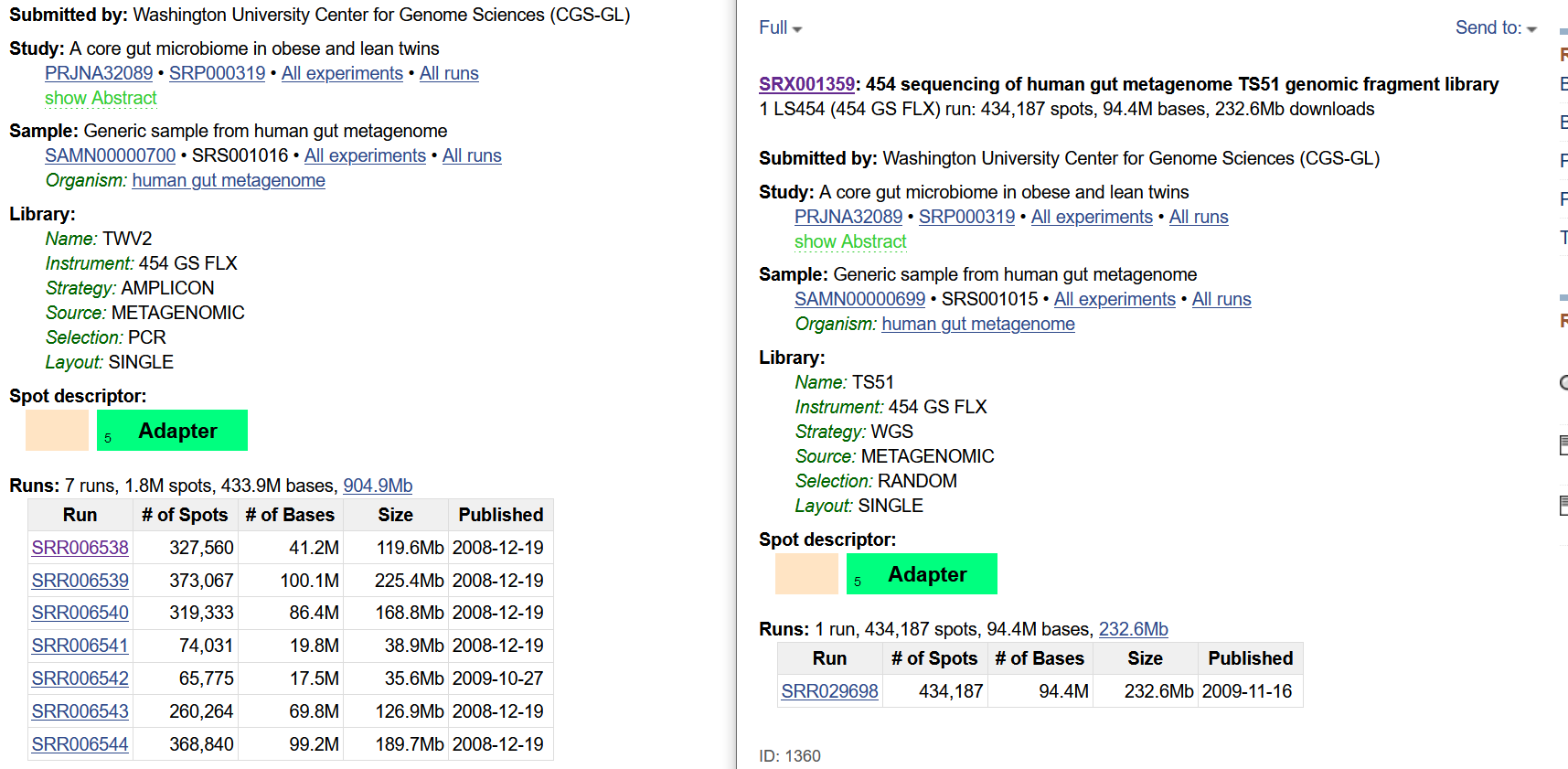
* ¿A que corresponden los datos que están en SRA?

**Respuesta:** Los datos que se encuentran almacenados en SRA corresponden a las secuencias obtenidas por la técnica de secuenciación 454 del microbioma intestinal analizado en el estudio.



Hay 2 tipos de datos que diferencia tienen?

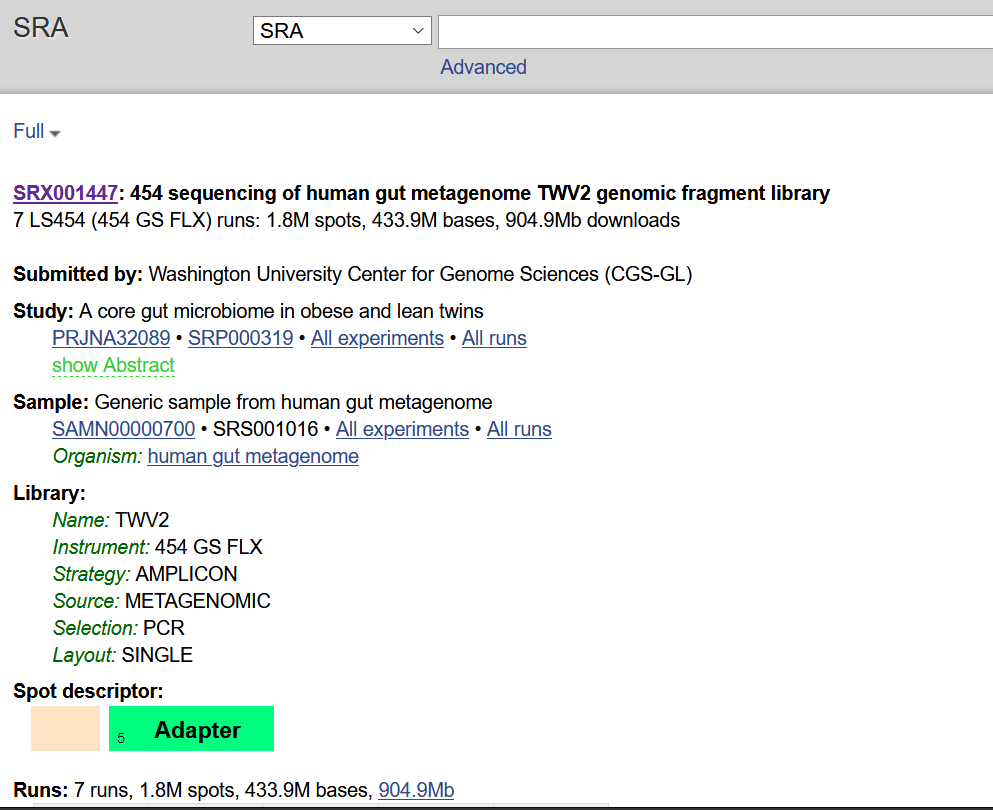
**Respuesta:** Al seleccionar los dos tipos de datos observamos que la principal diferencia se encuentra en la estrategia de secuenciación utilizada y en el método de selección empleado y finalmente existe una diferencia entre el número de corridas empleados por cada tipo de dato, esto debido a la diferencia de fecha en la que fueron realizados los experimentos, tal y como se muestra en la figura:



* Entrando al proyecto SRX001447. A que tipo de estrategia de secuenciación corresponde? Que instrumento se usó para generar estos datos? Cuantas corridas hacen parte de este experimento? ¿Y cuantas secuencias en total se generaron?

**Respuesta:** Luego de revisar el

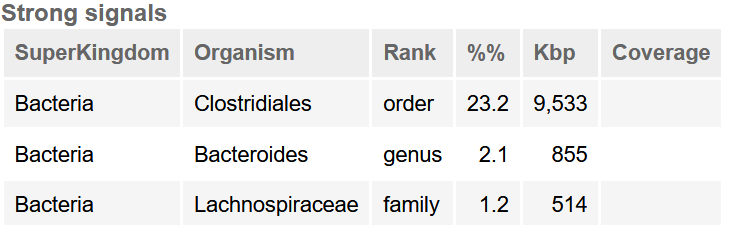
proyecto se observan las siguientes características:

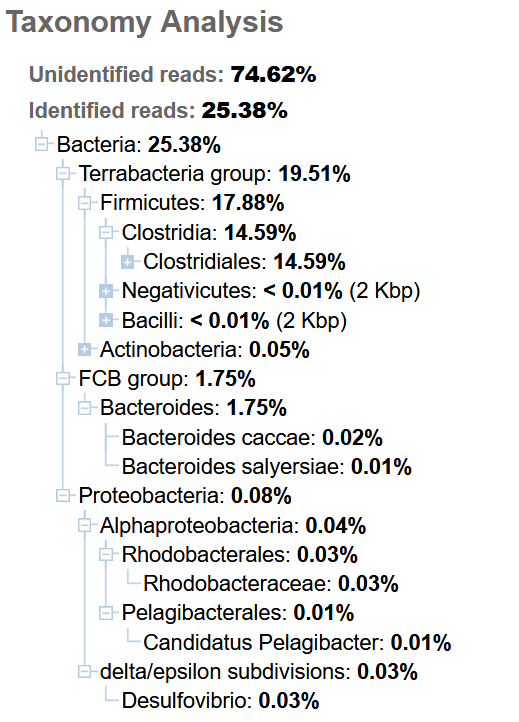


En donde se observa claramente que la plataforma que se utiliza para realizar la secuenciación 454 secuenciación, utilizando la herramienta 454 GS FLX para la obtención de los datos, para la obtención de los datos se realizan 7 corridas, generando 904.9Mb en información de las secuencias del estudio

* ¿Si entra a la corrida SRR006538, en la sección de análisis que información le da? Cuál es el phylum bacteriano más abundante?

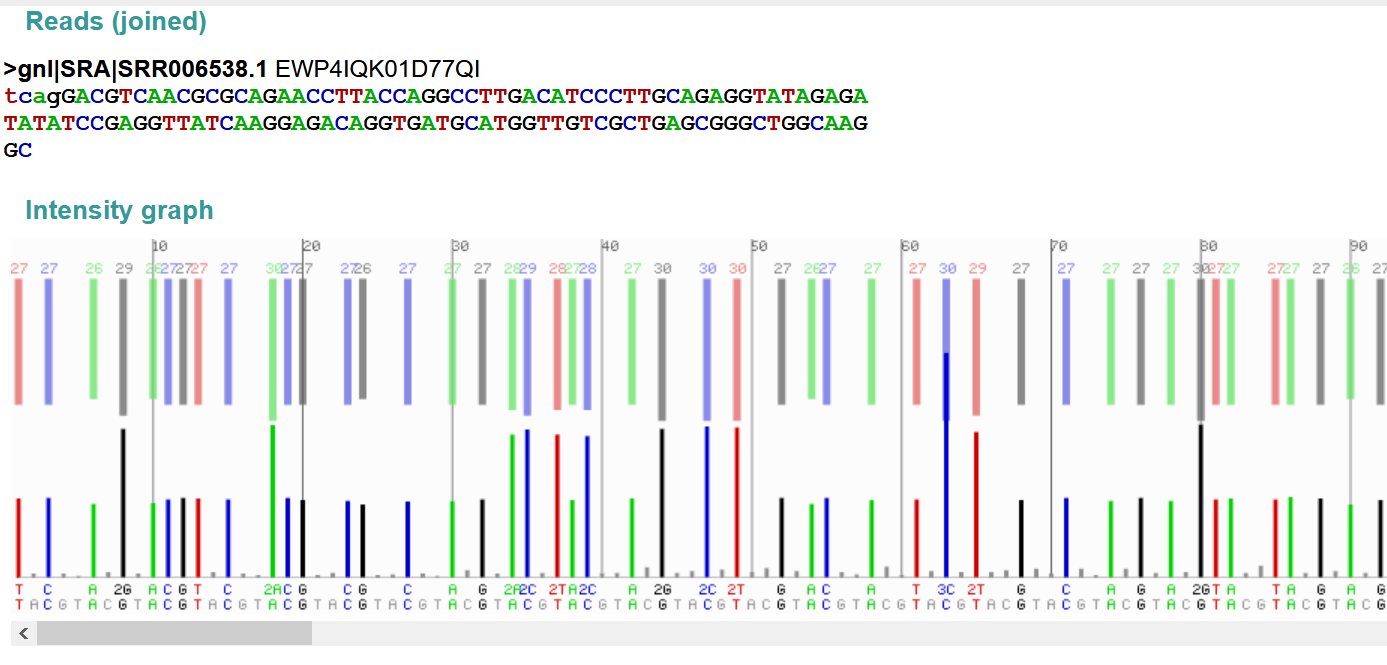
**Respuesta:** Luego de revisar la información de la corrida SRR006538 se obtiene la información mostrada en la figura que está adjunta, en ella se realiza una asignación del número conocido y desconocidos de reads que se obtienen durante la secuenciación, y dentro de los reads conocidos se realiza una identificación de los grupos de bacterias más abundantes que son el grupo de las terrabacterias, las esfingobacterias (FCB) y finalmente el grupo de las proteobacterias. Además, presenta una clasificación de las lecturas más fuertes encontrándose que las bacterias son los microorganismos más abundantes y que el orden de los Clostridiales tiene una mayor representatividad dentro de estos (23,2%).Por último, resulta evidente que el filum más abundante de todo el análisis es el fillum de los Firmicutes con una representatividad del 17.88%



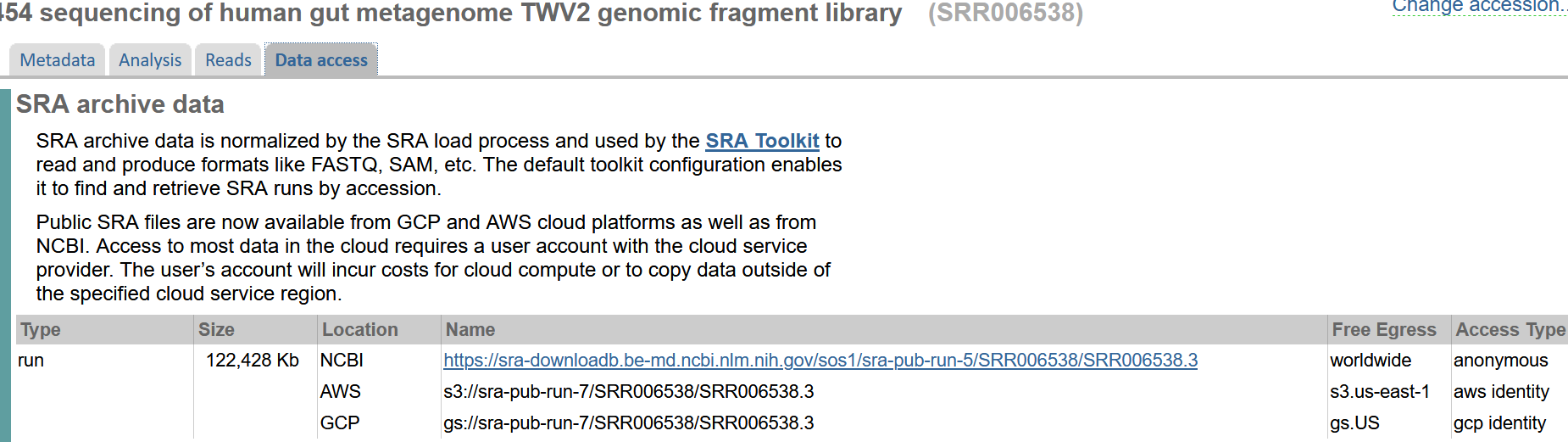


* ¿Que información le da la sección “reads” y la sección “Data Access”?

**Respuesta:** Luego de revisar la información de la sección *reads,* se obtiene una lista con todos los reads obtenidos y además se obtiene una secuencia de los reads juntos, además de esto se muestra un mapa en donde se asigna cada nucleótido

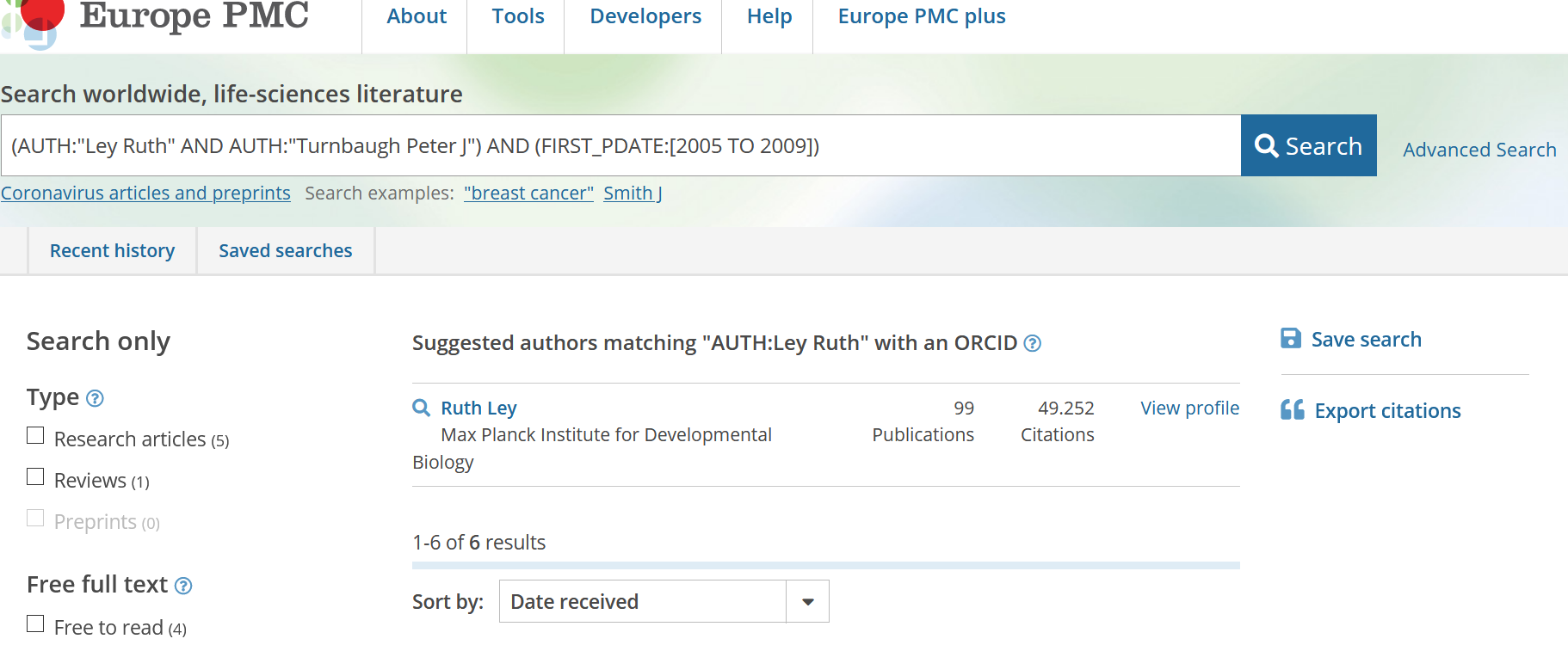


Al revisar la sección *Data Access,* es evidente que esta sección describe el peso que tienen las corridas, en donde se encuentran almancenados así como su nombre, y su tipo de acceso.



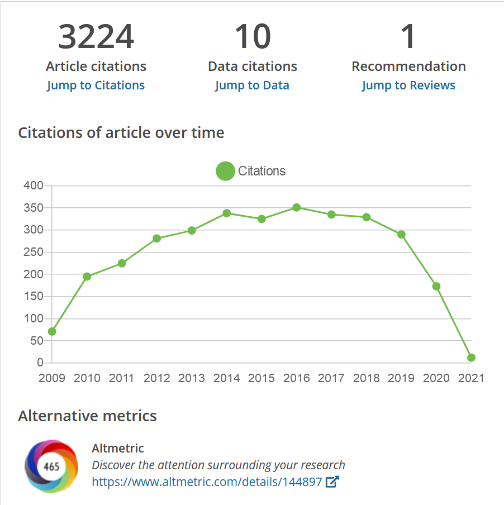
En el **EBI**, en <http://europepmc.org/>

Realice la misma búsqueda de artículos escritos por Ruth Ley y Peter J Turnbaugh entre 2005 y 2009. Como es la búsqueda? Cuantos resultados da? Son los mismos?

**Respuesta:** Al realizar la búsqueda se encontraron 6 artículos científicos entre sí, a la pregunta de si son los mismos que los encontrados en la búsqueda de los puntos anteriores usando la base de datos

* Seleccione nuevamente el mismo artículo. Cuantas citas obtiene? De un pantallazo de las citas en función del tiempo. ¿Si es diferente a los resultados del NCBI a que cree que se debe la diferencia?

**Respuesta:** Al realizar la búsqueda del artículo, se encontró que este se encuentra citado 3,224 veces, siendo el 2016 el año en el que más citas obtuvo.



Al realizar una comparación con los resultados obtenidos en la búsqueda en el NCBI (2,618 citaciones), resulta evidente que la cantidad de citaciones que presenta cada artículo es diferente, siendo el EBI la base de datos con mayor número de citaciones reportadas (3,224

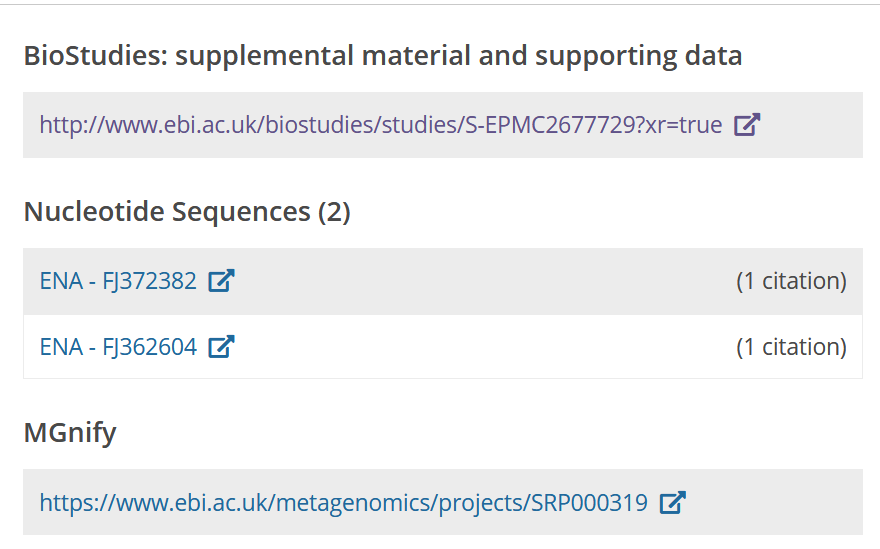
veces).

Una posible explicación de lo anterior se debe a que la base de datos del EBI utiliza un algoritmo que analiza conceptos de deep learning entrenado para clasificar las citas de acuerdo con tres categorías en específico, tal y como se describe en la siguiente figura:

* Aquellos autores realizan la citación aportando evidencia para soportar su validez
* Aquellos autores realizan la citación aportando evidencia para desaprobar o disputar su validez
* Aquellos autores que no brindan ningún soporte al momento de realizar la citación



* En la sección de datos hay vínculos nuevamente a Bio-estudios, secuencias de nucleótidos y Mgnify. A que corresponde cada uno de estos links?

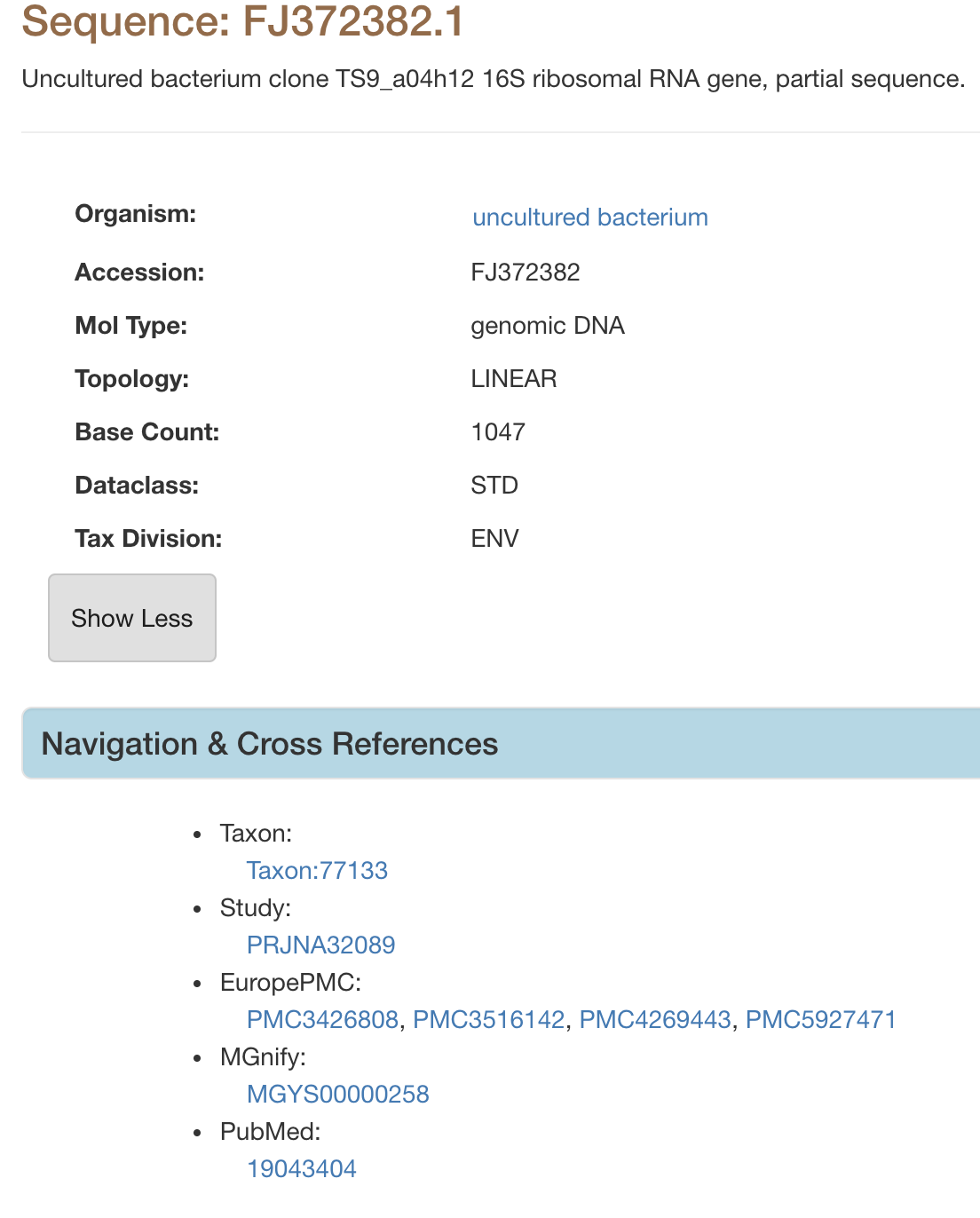


**Respuesta:** Al explorar la sección de datos, el resultado que se obtiene el apartado de cada una de las secciones es el siguiente:

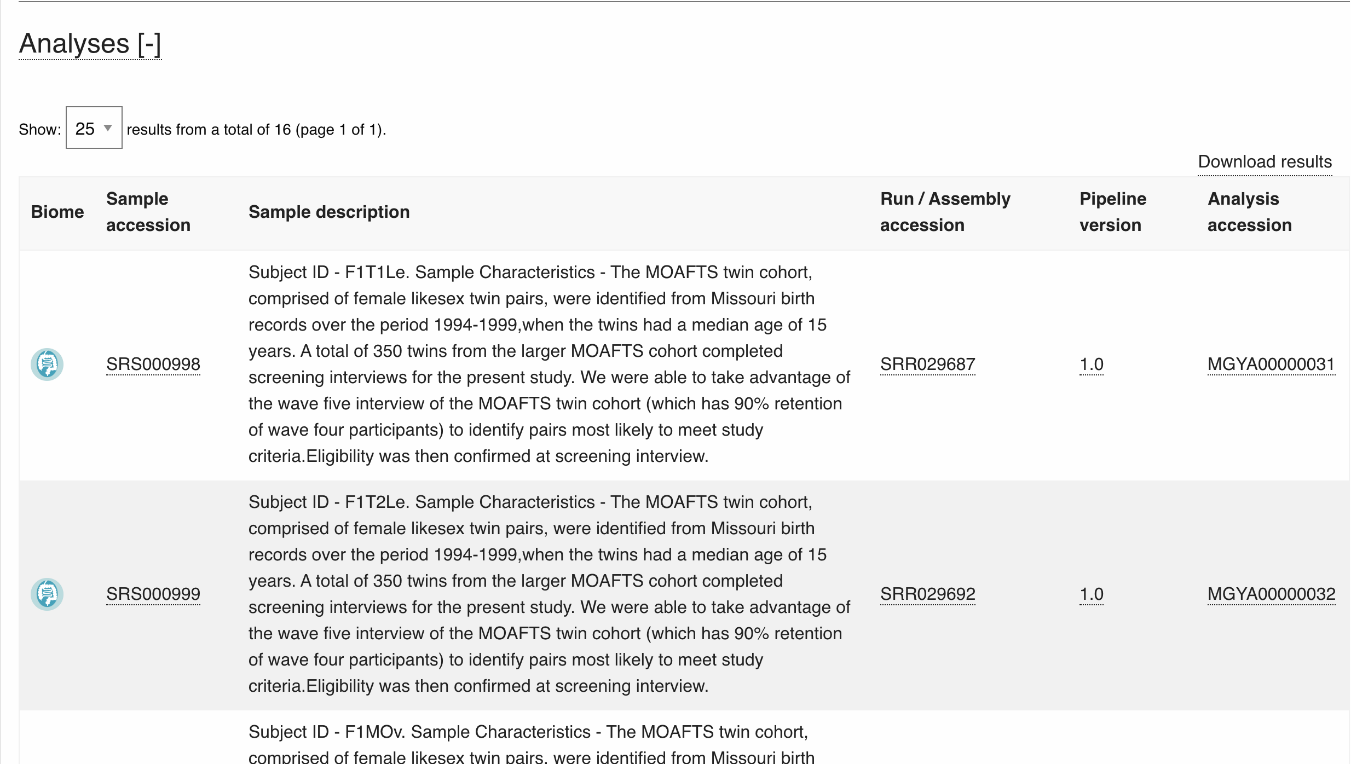
En el apartado de Bio-estudios, el link corresponde a información del estudio la cual incluye su financiación y sus datos de publicación.



En el apartado de secuencias de nucleótidos, el link corresponde a información particular de las secuencias, como su topología, el tipo de molécula a la que está asociada la secuencia y a las referencias cruzadas que la contienen.



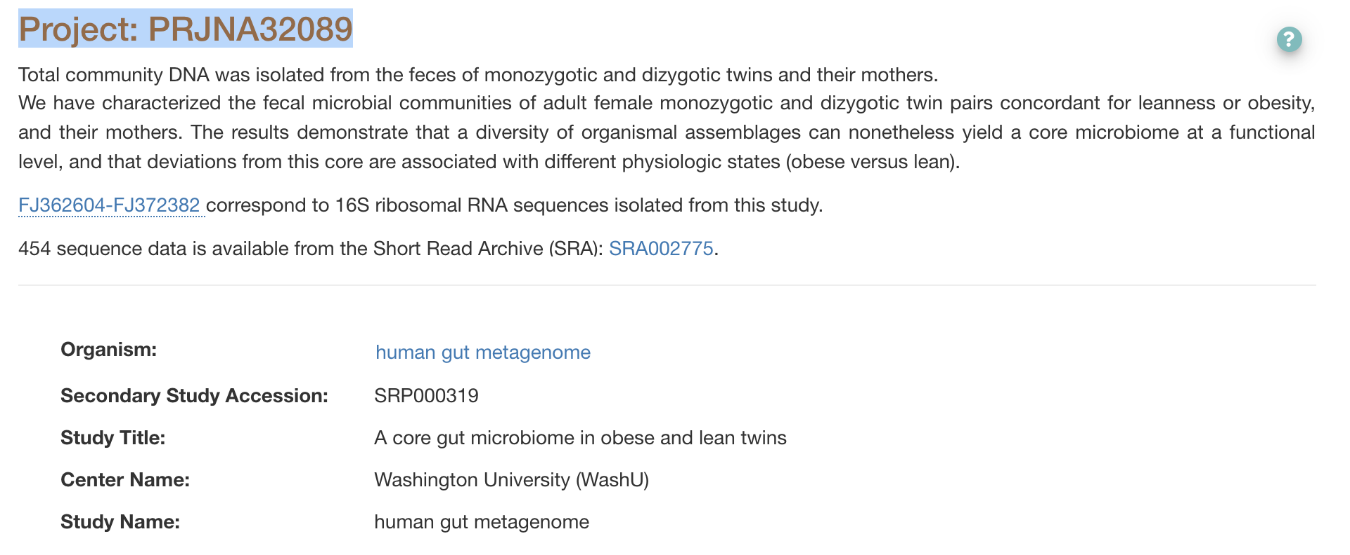
En el apartado de Mgnify, el link corresponde a una colección de diferentes análisis realizados sobre estudio.



* En el link de nucleótidos que lleva al ENA, hay referencia a un estudio. ¿El Accession number de este estudio es el mismo encontrado en el NCBI? A donde lleva ese vínculo?

**Respuesta**: Si, es el mismo *accession number* que encontramos en NCBI.

El vínculo nos dirige a un resumen donde encontramos información que varía desde los datos del estudio hasta la publicación del mismo.

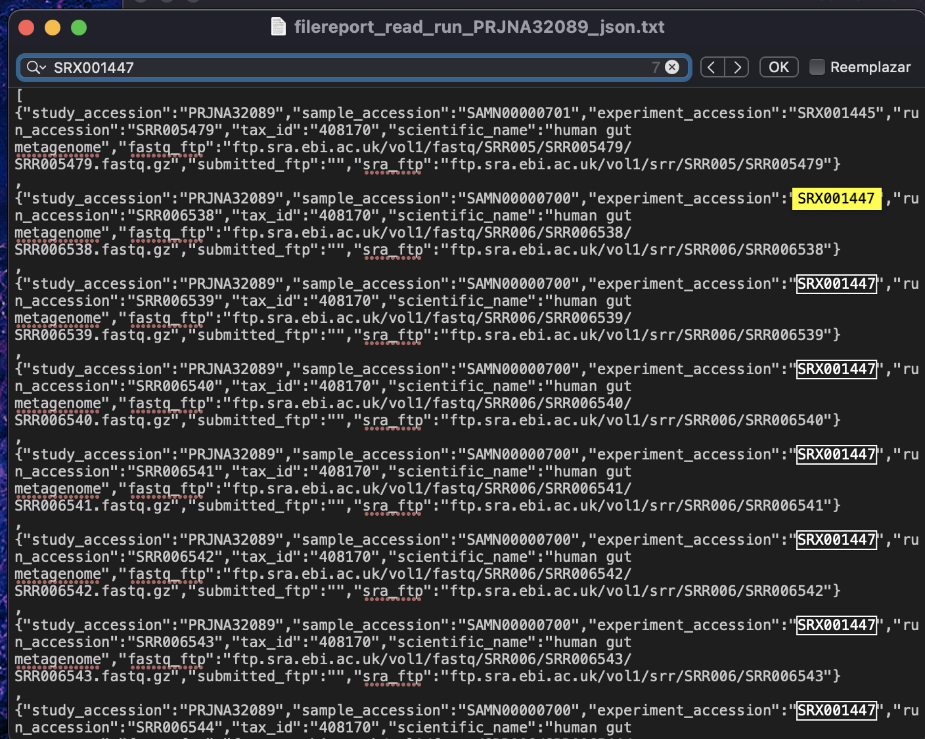
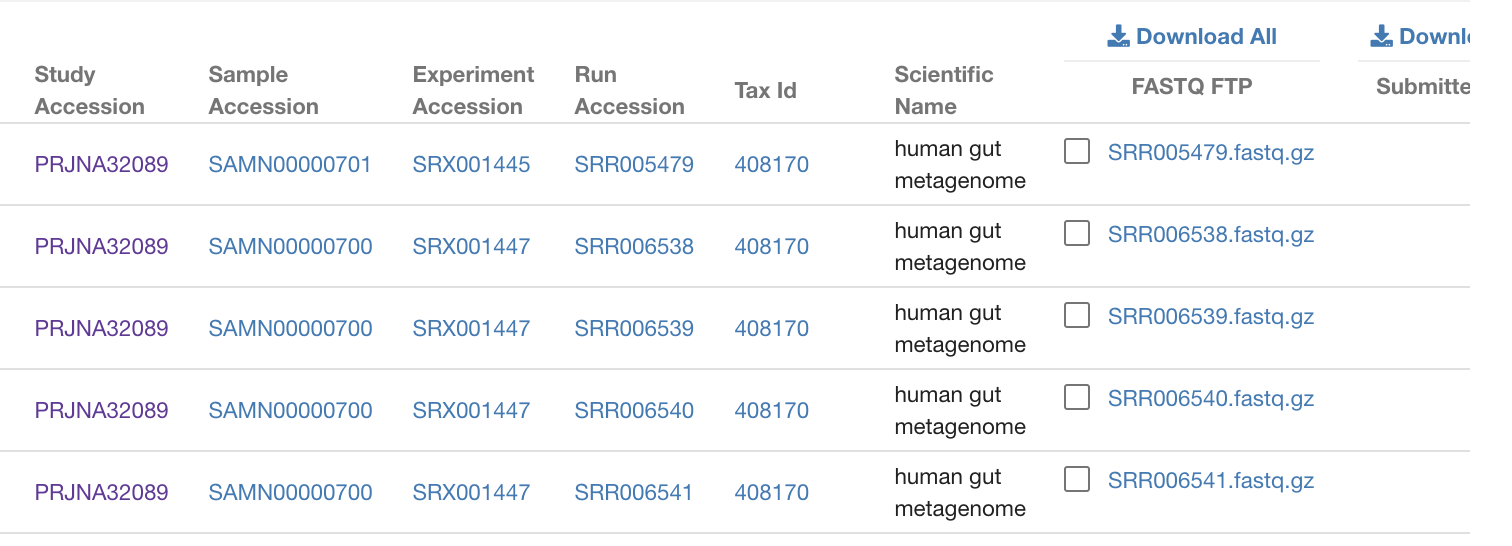


* Es posible encontrar las mismas corridas? El mismo experimento? Como se reconocen los códigos de acceso de las corridas o los experimentos? Permite ver los análisis o descargar los datos?

**Respuesta:** Si es posible encontrar las mismas corridas y el mismo experimento.

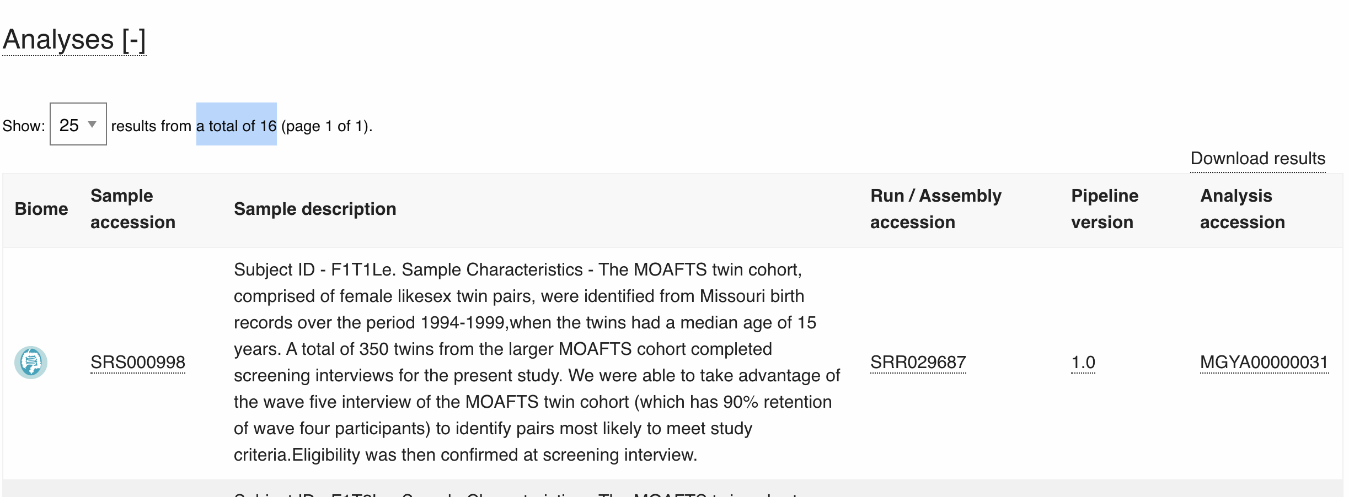
Los códigos de acceso de las corridas se reconocen porque empiezan con SRR, por otra parte, los códigos de acceso de los experimentos se reconocen porque empiezan por SRX. (nota: para facilitar el proceso de búsqueda descargamos los datos en un archivo JSON)

Permite descargar los datos, pero no permite ver los análisis.

* En el link a Mgnify cuantos análisis están disponibles para este estudio? Corresponden a los datos de secuenciación del gen 16S? o a shotgun metagenómico?

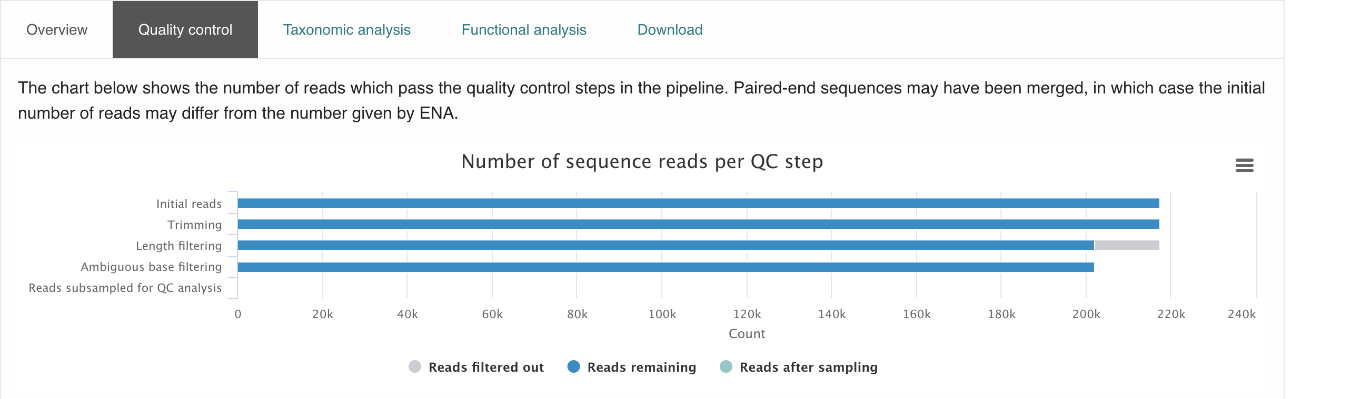
**Respuesta**:En Mgnify hay disponibles 16 análisis para este estudio. Los análisis corresponden a los datos de secuenciación del gen 16S.



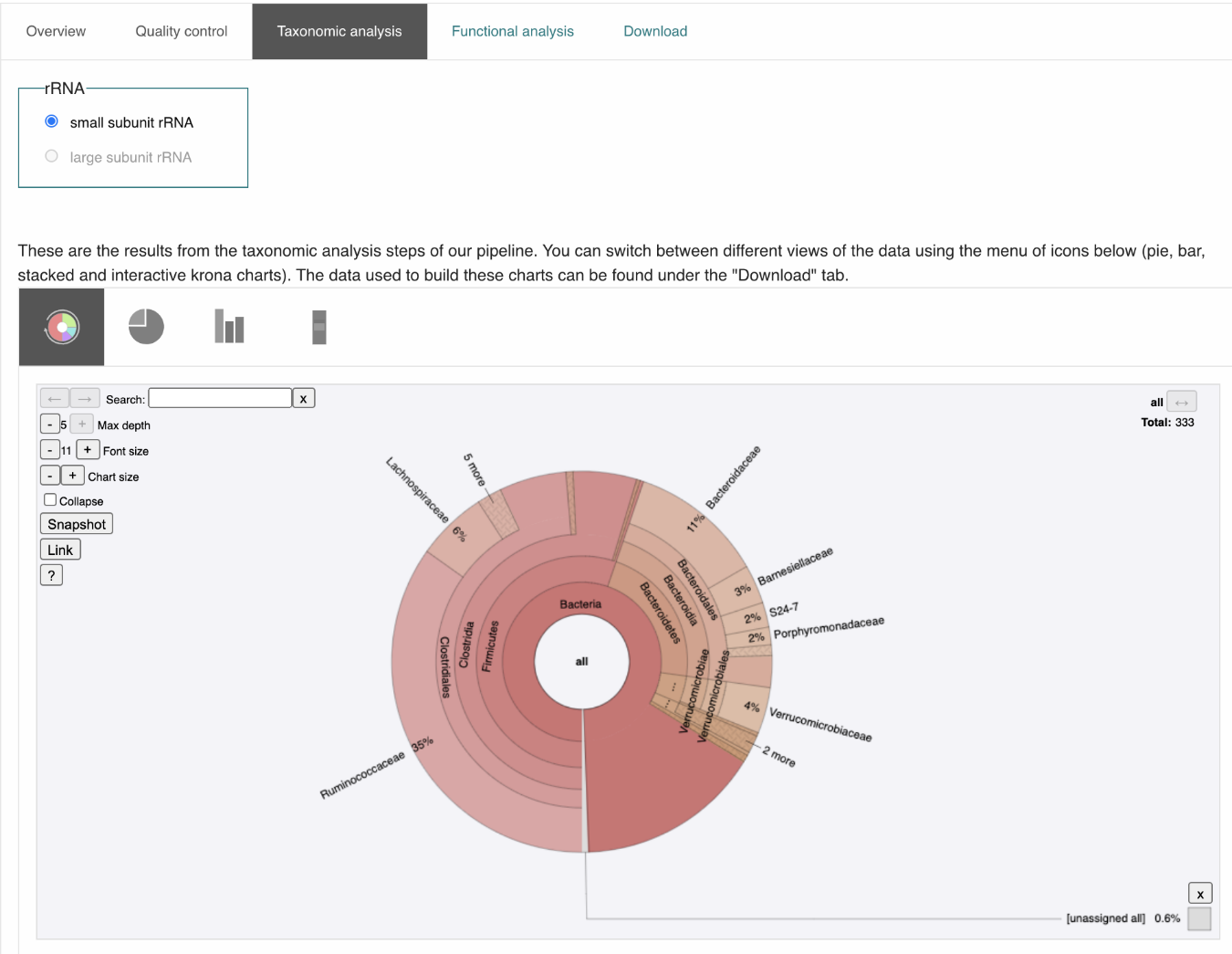
* + - * Entrando al primer análisis, correspondiente al SRR029687 ([MGYA00000031](https://www.ebi.ac.uk/metagenomics/analyses/MGYA00000031)). Que información le da las diferentes pestañas? (Quality Control, Taxonomic analysis, Functional analysis, Download).

**Respuesta:** La información que se ve en las diferentes pestañas es la siguiente:

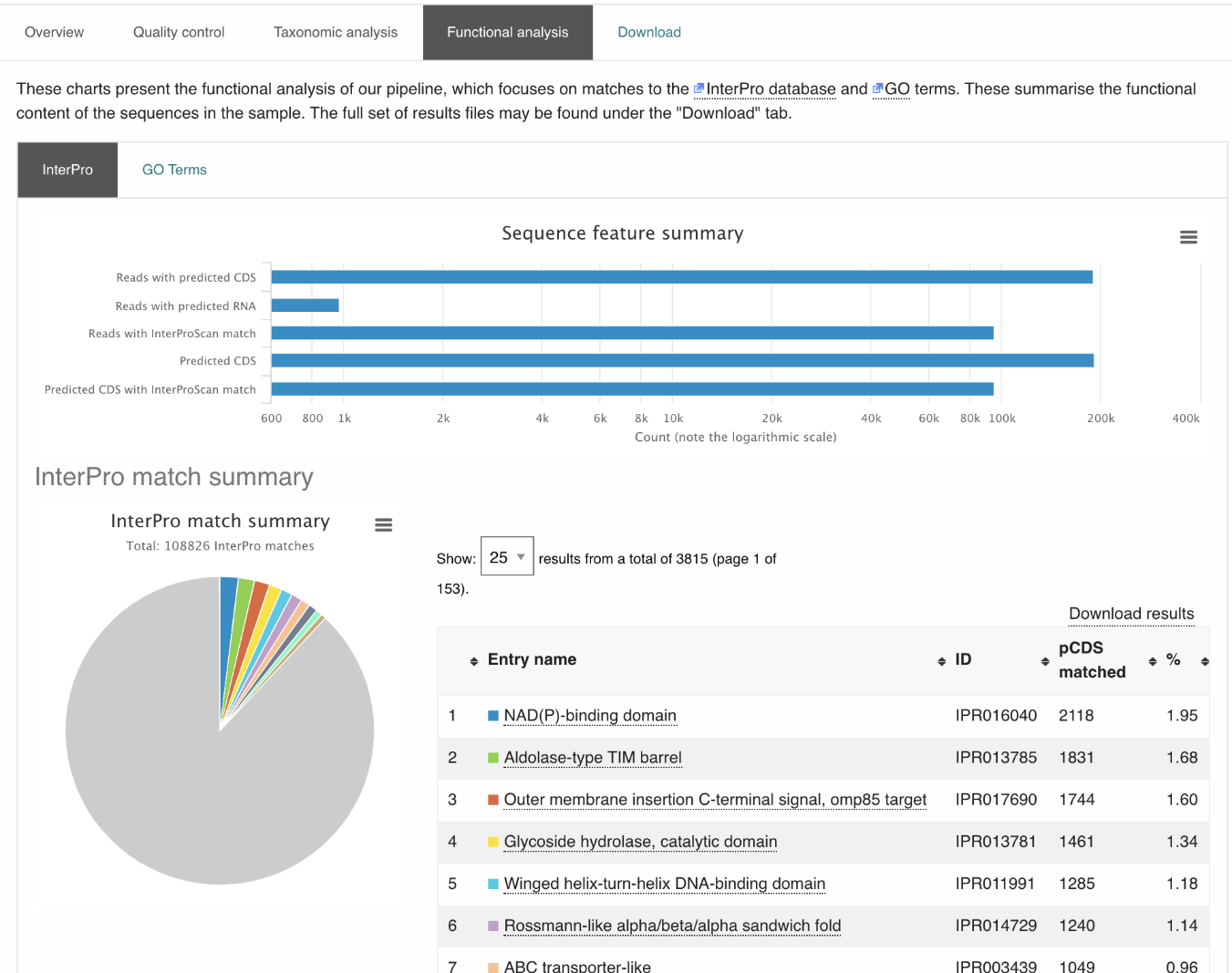
Quality Control: Numero de lecturas de secuencia por cada paso QC



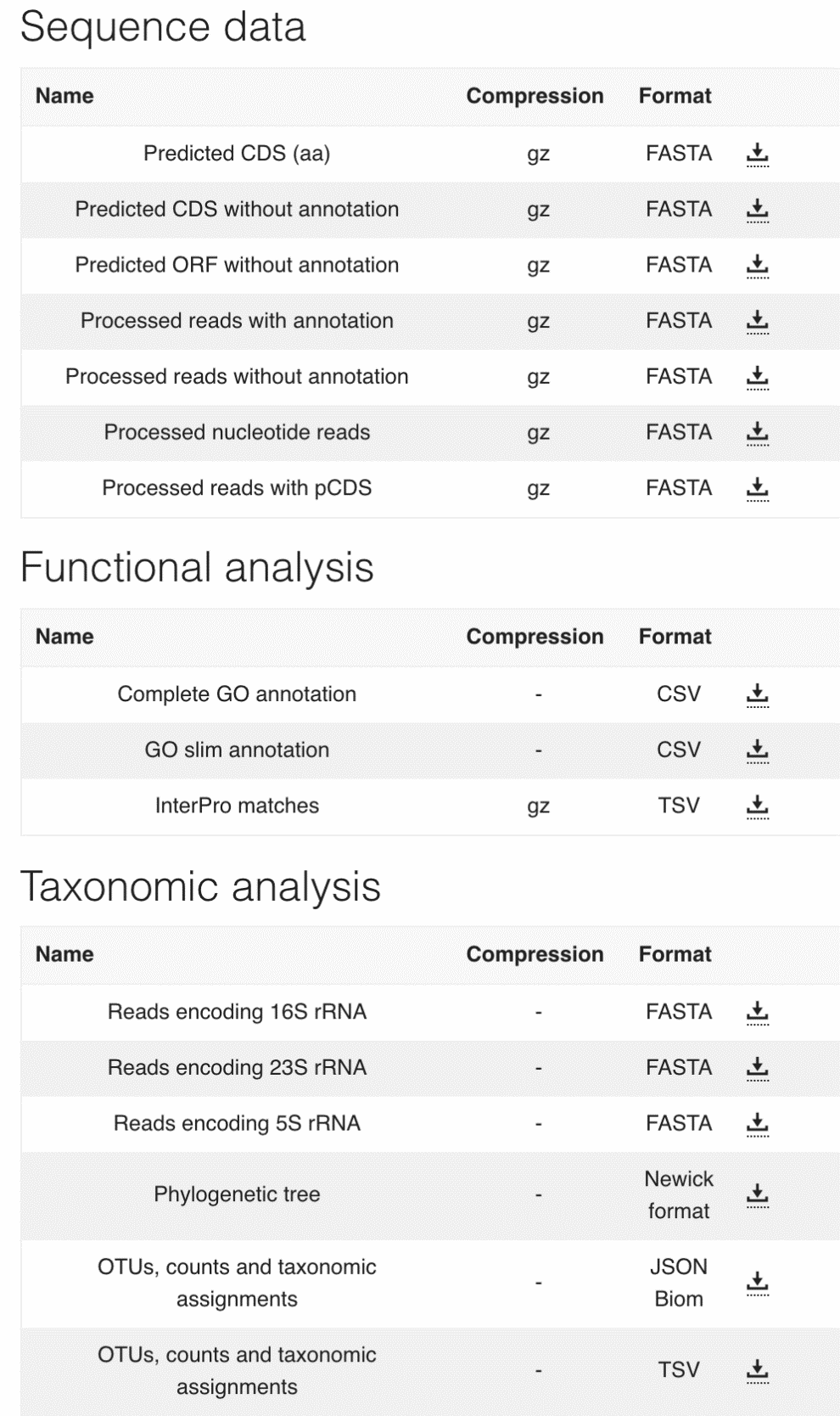
Taxonomic analysis:



Functional analysis:



Download: Se presenta la opción de descargar los datos presentados anteriormente (Sequence data, Funcional analysis y Taxonomic analysis) en diferentes formatos, algunos de los formatos son FASTA, CSV, TSV, JSON y Newick format.



* + - * ¿En su opinión personal considera más útiles los resultados obtenidos con el NCBI o con el EBI?

**Respuesta:** Consideramos más útiles los resultados obtenidos con el EBI, pues nos presentaron una perspectiva más completa sobre los diferentes motivos de citación del estudio; por otra parte, nos parecieron útiles las gráficas que podemos observar en los análisis disponibles, la cantidad de información que se puede obtener y la forma en la que esta plataforma permite acceder a los datos organizadamente facilitando a los investigadores lograr la reproducibilidad de los artículos presentados.

**Referencias**

Fernández-Altuna, M., Martínez del Prado, A., Arriarán Rodríguez, E., Gutiérrez Rayón, D., Toriz Castillo, H. and Lifshitz Guinzberg, A., 2016. Uso de los MeSH: una guía práctica. *Investigación en Educación Médica*, 5(20), pp.220-229.

scite help desk. 2021. How are citations classified?. [online] Available at: <https://help.scite.ai/en-us/article/how-are-citations-classified-1a9j78t/> [Accessed 5 February 2021].

Sequence Read Archive Submissions Staff. Understanding SRA Search Results. In: SRA Knowledge Base [Internet]. Bethesda (MD): National Center for Biotechnology Information (US); 2011-. Available from: https://www.ncbi.nlm.nih.gov/books/NBK56913/