Bitácora

Fecha 📅 25 de enero del 2024

# Hoy corroboré el acceso al cluster DNA, en el cual pude entrar y agregar un archivo plano llamado “saluditos.txt”.

# Posterior a ello, busque varios artículos publicados en Google Académico, con el fin de saber si tenían datos publicados de Pseudomonas y si cuales fueron las bases de datos usadas. Para ello a continuación, describiré lo entendido y en seguida encontrado de información

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Archivo | 1471-2164-14-271.pdf | | Articulo | **Comparative genomic analysis of four representative plant growth-promoting rhizobacteria in Pseudomonas** | |
| Observaciones | | | *25 de enero del 2024*Al parecer que con lo explica en el resumen, el articulo entiendo que hay cepas que tienen distintos nichos, aunque sea la misma especie. A que voy con esto, dentro de P. aeruginosa, hay bacterias que son patógenas, pero también hay bacterias que son promotoras de crecimiento de las plantas. Observo que en los resultados hablan sobre los grupos ortólogos y secuencias codificantes (CDSs). Ahora, ¿qué herramientas usaron para el análisis y de donde sacaron los datos? Perfecto. Al leer en los métodos, se mencionó que el genoma que secuenciaron, lo colocaron en IMG database (<https://img.jgi.doe.gov/>). A continuación, encontramos que muchos quedan en “Draft” o sea que solo se ha quedado a nivel de borrador, pero no como genoma completo. Para esto otra observación, es que veo más de otro tipo de pseudomonas, las cuales son **Halopseudomonas, Pseudomonadota**  En otras observaciones, al ver que los datos cada uno tiene un nombre de muestra, podemos ver que las accesiones de los datos se dirigen a NCBI. Así en la tabla general de la búsqueda ya hecha por el organismo te da en este orden los datos (Select, Domain, Sequencing Status, Study Name, Genome Name/ Sample Name, Sequencing Center, IMG Genome ID, Species). Donde Study Name te da la descripción de donde fue extraído y que cepa es.  Tomando en cuenta que la gran mayoría son “Drafts” y no secuenciaciones finalizadas. Es de preguntar si pueden funcionar.  Ahora revisando que se refiere los demás nombres taxonómicos, en <https://www.ncbi.nlm.nih.gov/Taxonomy/CommonTree/wwwcmt.cgi>  --  Este es el esquema, donde los marcados en **negritas** son los respectivos  superkingdom, kingdom, phylum, class, order, family, **genus**, species  cellular organisms; Bacteria; Pseudomonadota; Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; **Pseudomonas**  Y este es el esquema que sale cuando se compara:   |  | | --- | | [**Pseudomonadota**](https://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?mode=Tree&id=1224) (phylum) | | branch[Gammaproteobacteria](https://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?mode=Tree&id=1236) (class) | | emptybranch[Pseudomonadales](https://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?mode=Tree&id=72274) (order) | | emptyemptyCollapse[Pseudomonadaceae](https://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?mode=Tree&id=135621) (family) | | emptyemptyemptybar with branch[**Halopseudomonas**](https://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?mode=Info&id=2901189) | | (genus 1)  emptyemptyemptybranch[**Pseudomonas**](https://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?mode=Info&id=286) (genus 2) | | | |
| Fecha de primera consulta | | 25/01/2024 | Bases de datos encontradas e información de interés | | IMG database <https://img.jgi.doe.gov/> |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Archivo | srep38699.pdf | | Articulo | **Comparison of 432 Pseudomonas strains through integration of genomic, functional, metabolic and expression data** | |
| Observaciones | | | *26 de enero del 2024*Partimos con leer el articulo para hacer la búsqueda de la base de datos, Ahora que leí el articulo, encontré que tienen data extra, de la cual se averiguara. Fue descargado los datos suplementarios que tiene como zip name: 41598\_2016\_BFsrep38699\_MOESM1\_ESM y fue almacenada en una carpeta con el nombre del PDF, dentro de la carpeta Data Download que así mismo se encuentra en la carpeta donde se están consultado estos datos. **Veo que, al parecer el ZIP, solo tiene archivo XML y \_MATLAB y \_MODELS, pero yo al entrar con mi cuenta de la universidad, no me deja y no me llegan correos de confirmación, entonces intentare otro día Se intentara otro dia por eso en rojo.**  Hablando de los métodos. Por lo visto se descargaron 58 genomas circulares y 374 borradores del genoma del genero de las Pseudomonas, desde la base de datos de GenBank, por otra parte, las anotaciones, pueden funcionar para un futuro, las cuales fueron descargadas de RAST.  Por otra parte, menciona el articulo que pone toda la información a disposición en esta pagina, <http://semantics.systemsbiology.nl>. La cual te redirige a este link [Systems and Synthetic Biology - WUR](https://www.wur.nl/nl/Onderzoek-Resultaten/Leerstoelgroepen/Agrotechnologie-en-Voedselwetenschappen/Laboratory-of-Systems-and-Synthetic-Biology.htm) pero por lo visto, no carga algún sitio en donde se puedan cargar los datos, se pueden consultar ese y otras publicaciones que han hecho. Pero no tiene directamente los datos.  Por otra parte, no se encuentra información de la cual hable sobre la patogenicidad de las Pseudomonas, por lo que puede servir este documento para hacer una posterior consulta con respecto a las, pero si menciona sobre las secuencias fueron sacadas de NCBI, por lo que ahora redirigire la búsqueda a NCBI, presionar este [link](#enero26_2024_1) para continuar con la observación de este dia | | |
| Fecha de primera consulta | | 26/01/2024 | Bases de datos encontradas e información de interés | | <https://www.ncbi.nlm.nih.gov/genome> |

Link de **Comparison of 432 Pseudomonas strains… a la busqueda de NCBI,** [**presionar**](#enero26_2024_2_marcoexplicativo) **para ir a donde esta la anterior explicación.**

Fecha 📅 26 de enero del 2024

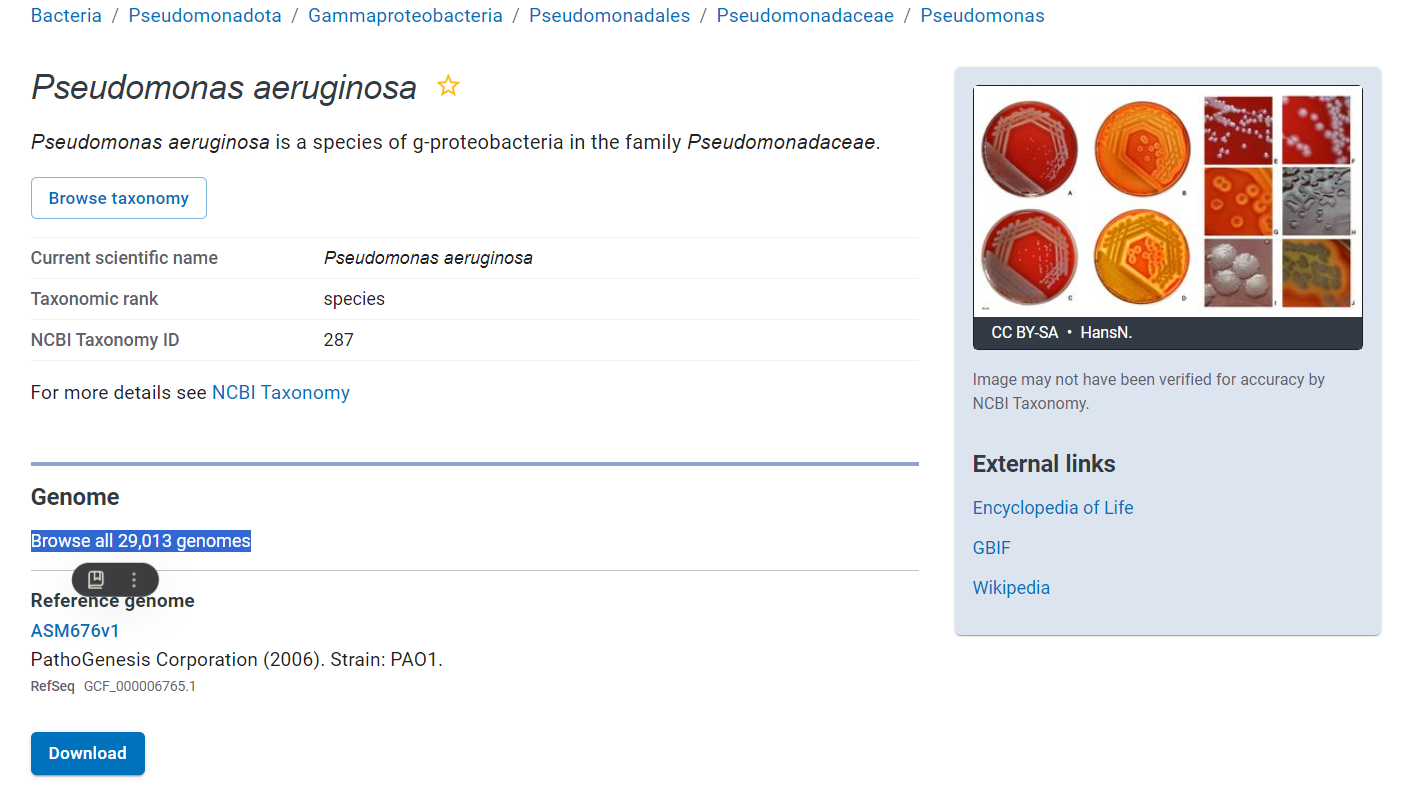
Se busco como “**pathogen pseudomonas genoma”** y arrojo, muchas especies de Pseudomonas, de las cuales algunas si son claramente descritas como patógenas y provenientes de determinados organismos o acciones y otras que no especifican de donde fueron aislados o si fueron patógenas.

Para esto se muestra esta tabla elaborada de información recopilada de <https://www.ncbi.nlm.nih.gov/genome/?term=pathogen+pseudomonas+genome>

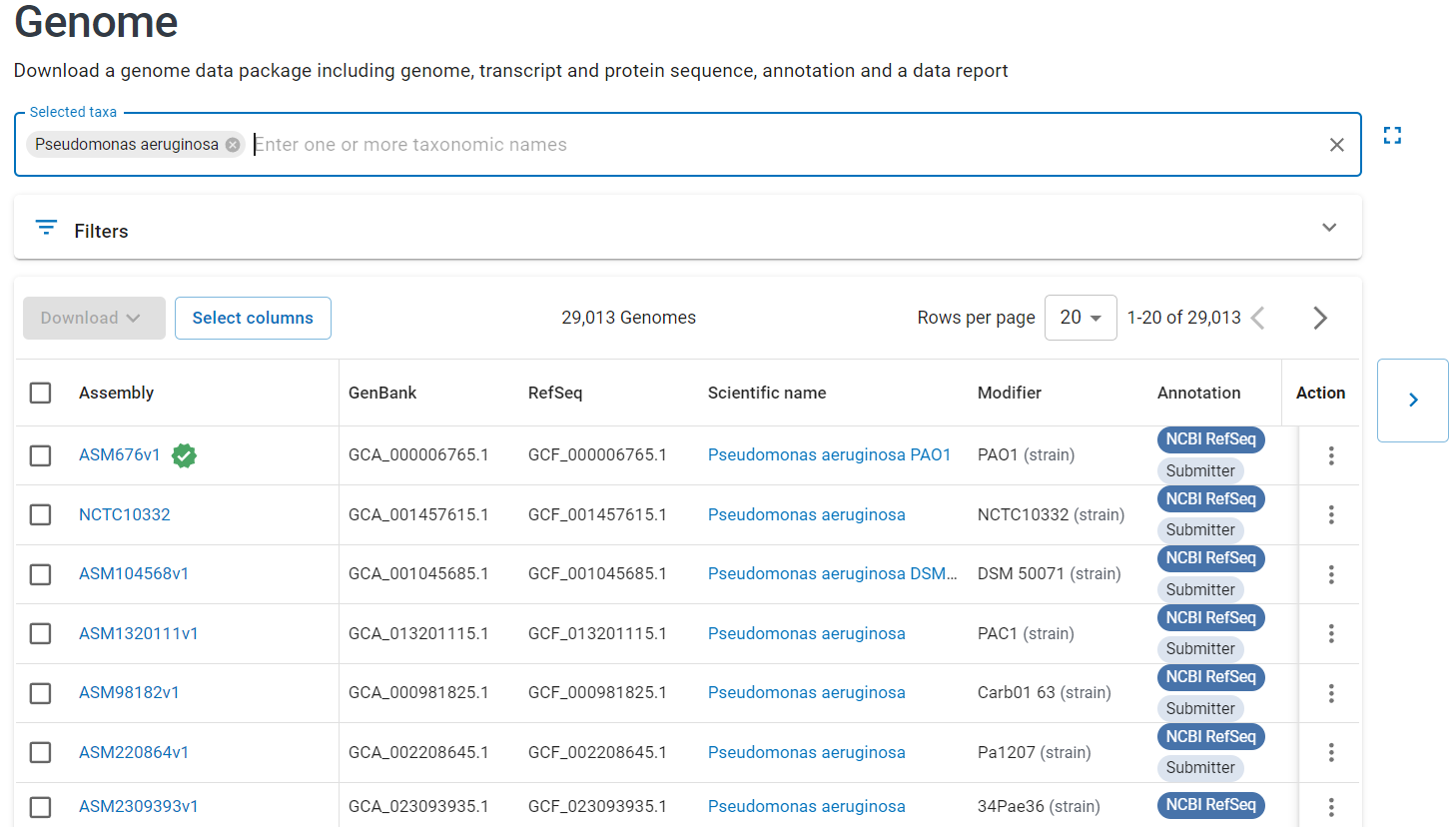
Que es la recopilada de las 13 especies que da la busqueda

|  |  |
| --- | --- |
| 1. [**Pseudomonas** aeruginosa](https://www.ncbi.nlm.nih.gov/genome/187)  Opportunistic **pathogen**  Kingdom:  Bacteria  ;  Subgroup:  Gammaproteobacteria  Sequence data: **genome** assemblies:  8880  Chromosome:  1  ;  Plasmids:  5  Date:  2000/05/16  ID:  187  Select item 185  2. [**Pseudomonas** syringae](https://www.ncbi.nlm.nih.gov/genome/185)  Plant **pathogen**  Kingdom:  Bacteria  ;  Subgroup:  Gammaproteobacteria  Sequence data: **genome** assemblies:  696  Chromosome:  1  ;  Plasmids:  5  Date:  2005/05/12  ID:  185  Select item 174  3. [**Pseudomonas** putida](https://www.ncbi.nlm.nih.gov/genome/174)  Common environmental bacterium  Kingdom:  Bacteria  ;  Subgroup:  Gammaproteobacteria  Sequence data: **genome** assemblies:  265  Chromosome:  1  ;  Plasmids:  4  Date:  2002/02/14  ID:  174  Select item 2506 | 4. [**Pseudomonas** amygdali](https://www.ncbi.nlm.nih.gov/genome/2506)  Produces a hyperplastic canker in the almond tree  Kingdom:  Bacteria  ;  Subgroup:  Gammaproteobacteria  Sequence data: **genome** assemblies:  114  Chromosome:  1  ;  Plasmids:  6  Date:  2009/03/02  ID:  2506  Select item 12176  5. [**Pseudomonas** oleovorans](https://www.ncbi.nlm.nih.gov/genome/12176)  **Pseudomonas** oleovorans overview  Kingdom:  Bacteria  ;  Subgroup:  Gammaproteobacteria  Sequence data: **genome** assemblies:  32  Chromosome:  1  Date:  2013/12/17  ID:  12176  6. [**Pseudomonas** viridiflava](https://www.ncbi.nlm.nih.gov/genome/11288)  **Pseudomonas** viridiflava overview  Kingdom:  Bacteria  ;  Subgroup:  Gammaproteobacteria  Sequence data: **genome** assemblies:  1504  Chromosome:  1  ;  Plasmids:  1  Date:  2012/10/22  ID:  11288 |

Para ello cuando se realiza la búsqueda de los genomas de especies que ya se sabe que son patógenas, como por ejemplo



Se encuentra un link para el cual se encuentran muchos mas genomas de la misma especie. Asi cuando se ingresa, aparece la siguiente información:



Con esto, podría yo suponer que todas las cepas que son reportadas, son patógenas. Pero quedan dos cuestiones en frente, una es que ¿cuál es archivo que contiene el genoma? Dos, ¿es bueno suponer que todos estos genomas son cepas patógenas?

Para esto me hago la pregunta, ¿que si al consultar en la base de datos IMG, ahí se podrá distinguir fácilmente si son o no patógenas y cual es el genoma?

Para esto busqué en la base de datos y encontré que fueron extraídos como patógenos y tienen información que puede ser de ensamble, ref y etc., me hace falta averiguar que es que cada archivo y si los datos me pueden aportar genomas útiles para el uso del análisis

Fecha 📅 27 de enero del 2024

Para este día, se visualizó la página <https://sites.google.com/a/ciad.mx/bioinformatica/home/genomica/obtencin-de-genomas> (**explica cómo obtener varias secuencias de un script**) y conocer los tipos de archivos que se catalogan dentro de NCBI <https://www.ncbi.nlm.nih.gov/home/documentation/>, para lo cua pasamos a este link <https://www.ncbi.nlm.nih.gov/genome/doc/ftpfaq/> para hacer la denotación de los distintoos tipos de archivos existentes.

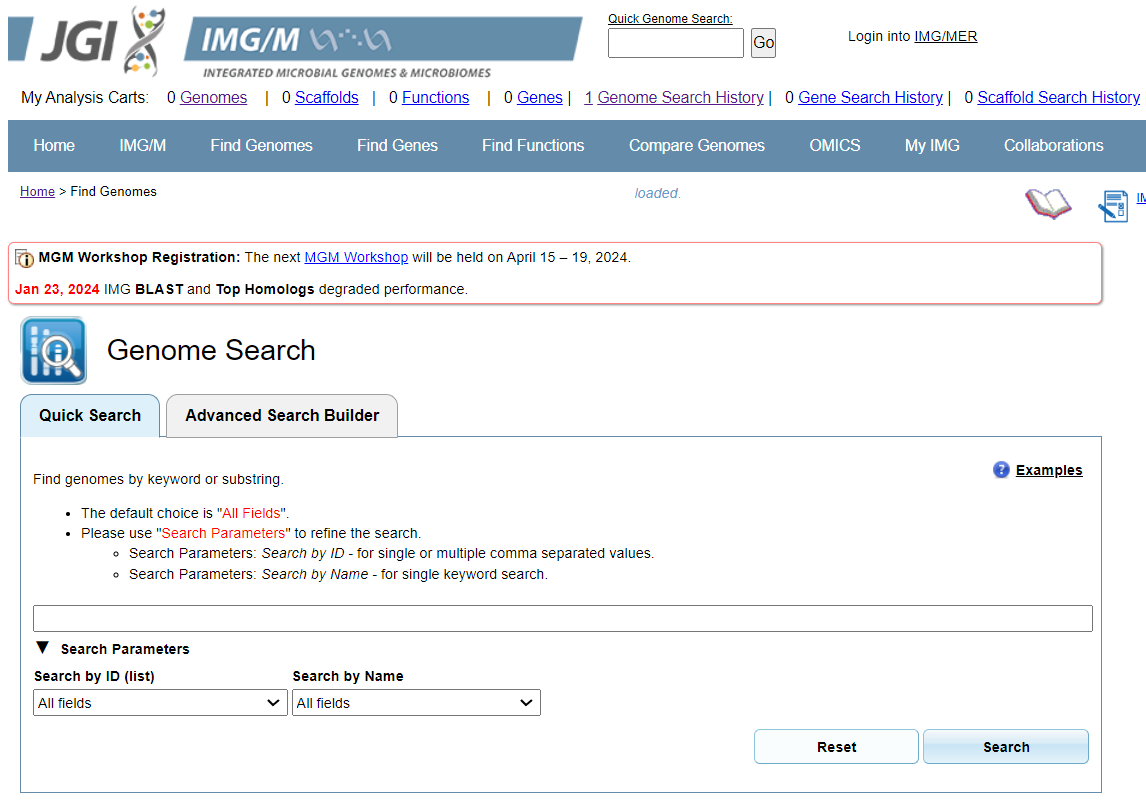
Resumen

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Los datos de los genomas de GenBank y los genomas de RefSeq están colocadas de lugres distintos | | | | Fecha de modificación |
| Tipo de datos | Tipo de archivo | Type Accession | Descripcion | 27/01/2024 |
| GenBank | gbcon\*.seq | GCA\_xxxxxxxxx.x | No contiene datos de secuencia en todo el archivo, el cual contiene instrucciones para la construcción objetos de grandes escalas. Ademas puede contener informacion de los genomas  <https://ftp.ncbi.nlm.nih.gov/genbank/README.genbank> | 27/01/2024 |
| Genome | .fna o .fasta … | GFC\_xxxxxxxxx.x | Contiene el titulo (nombre del organismo o división por chromosome) y contiene la secuencia. | 27/01/2024 |

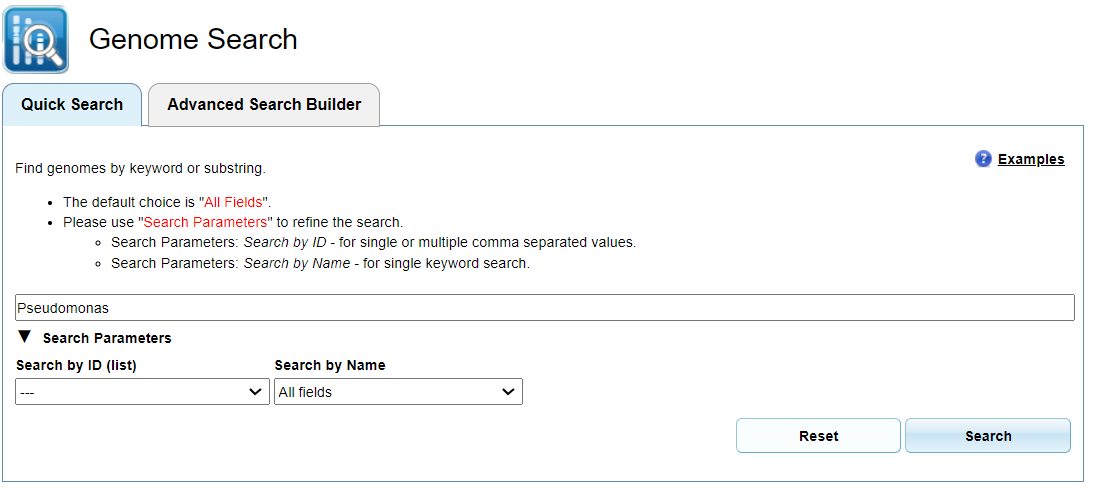
Para continuar ahora, procederemos a recopilar las accesiones para hacer la descarga de los archivos en el cluster

Si no antes, para esto, analizamos (como ejemplo).

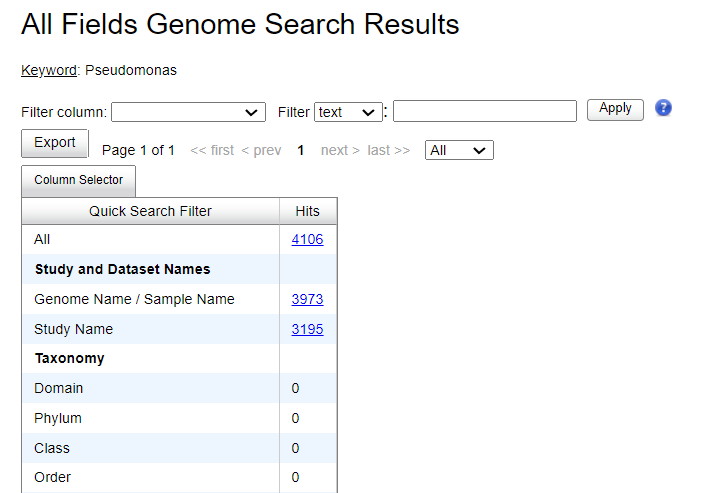
En la búsqueda de secuencias, entramos a IMG databases, y en la sección Find & Analyze > Genomes & Metagenomes



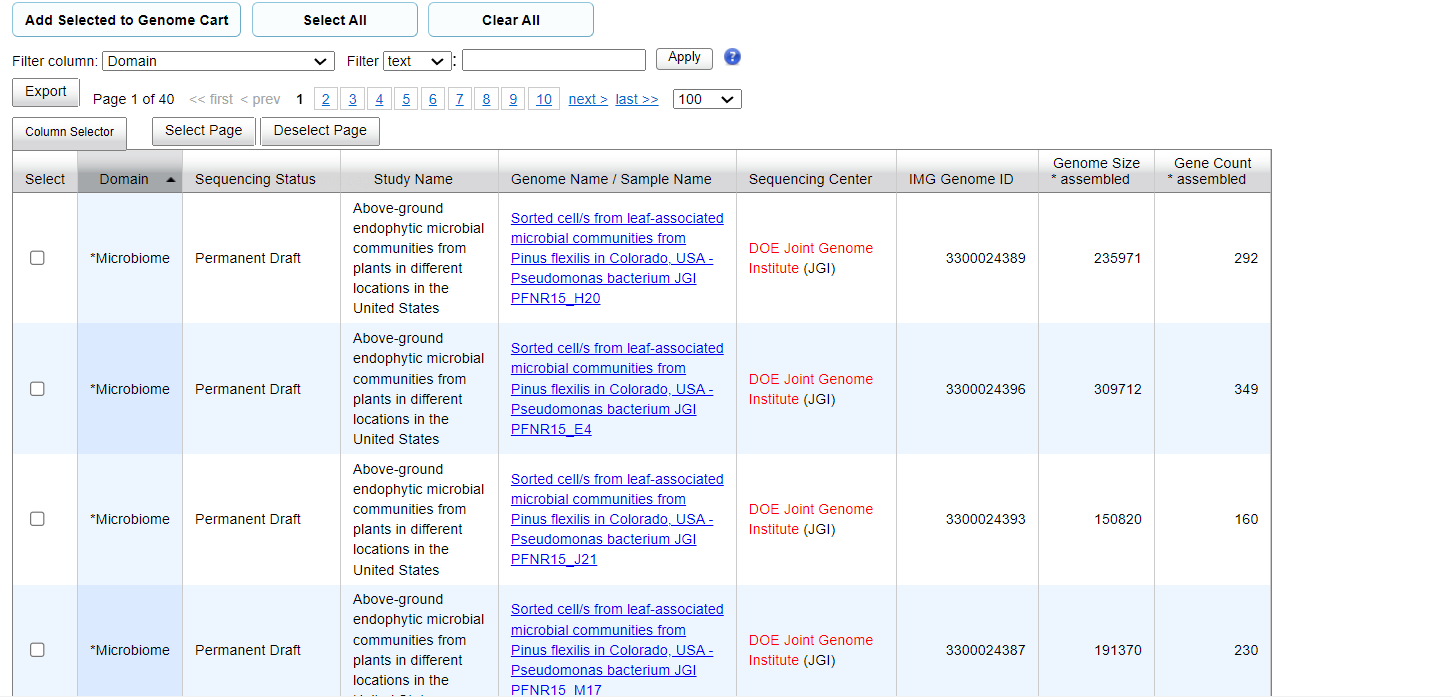
Damos en el campo de búsqueda y escribimos Pseudomonas



Al buscar arroja una tabla con información de los genomas



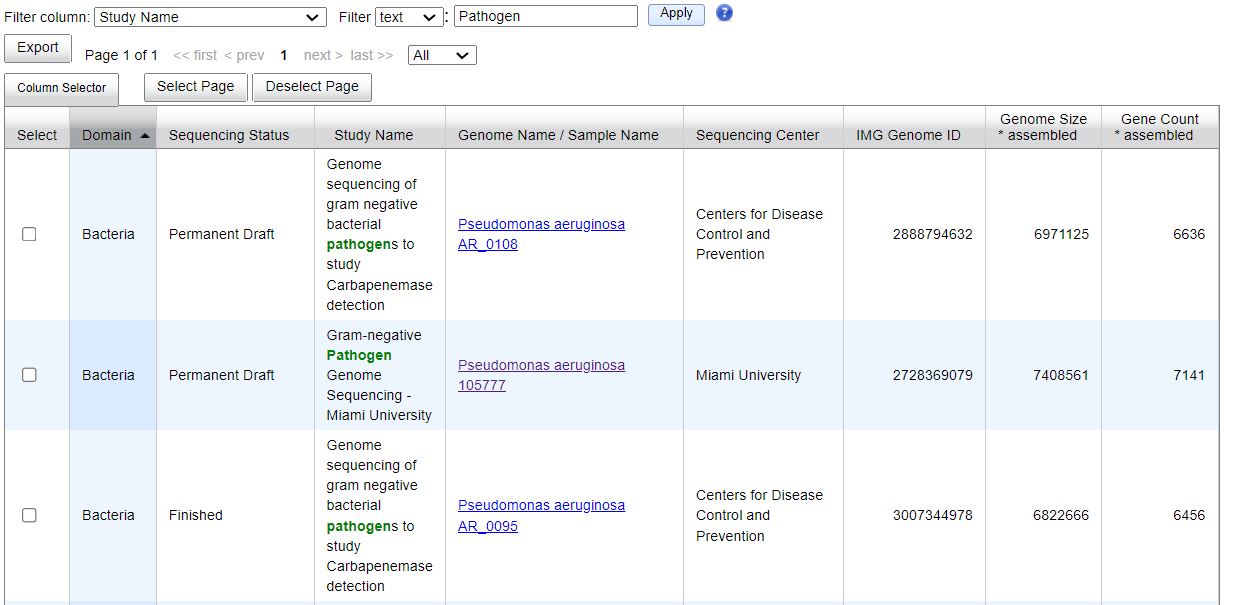
Para ello presionamos en el link (numero 3973) de Genome Name / Sample Name y para que ello nos de todos los datos



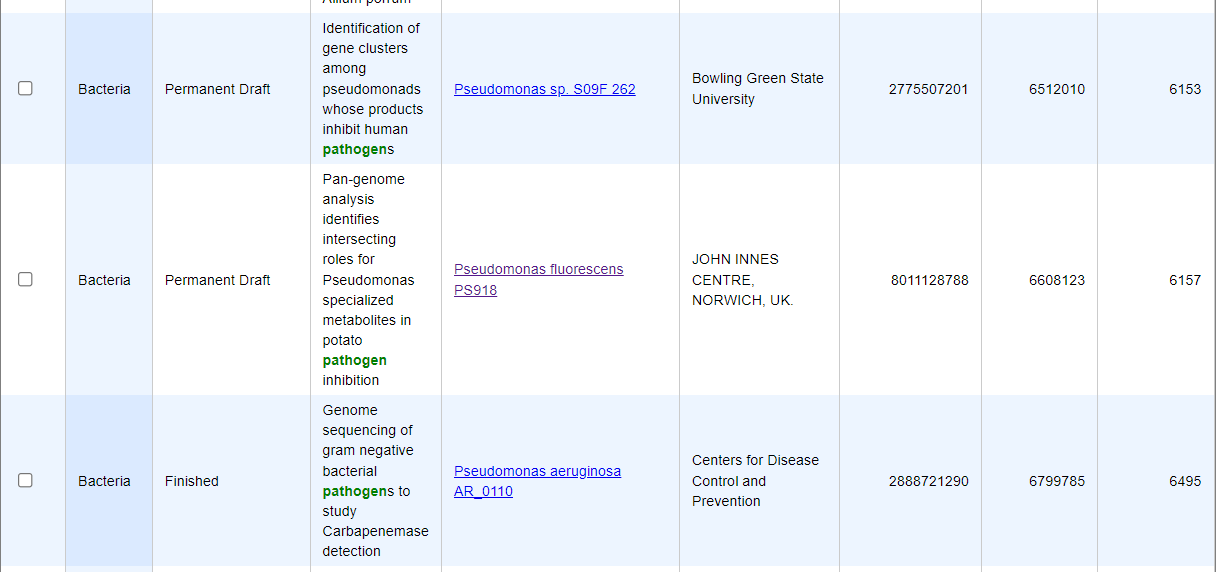
Configuramos estos parámetros y buscamos



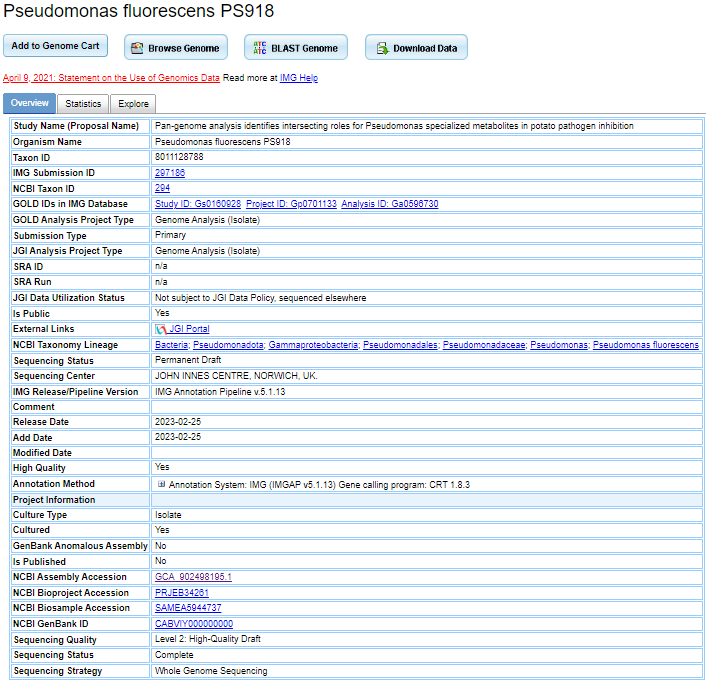
En lo que nos da datos sobre algunos estudios que tienen relación con patógenos, pero que no necesariamente aclara que las pseudomonas son las patógenas, sino que incluso son hospederas y combaten a otros patógenos.



Pero en consideración, buscamos como ejemplo este organismo (Pseudomonas fluorescens PS918) que como dice la descripción parece que las Pseudomonas juegan un rol importante en el control de patógenos.



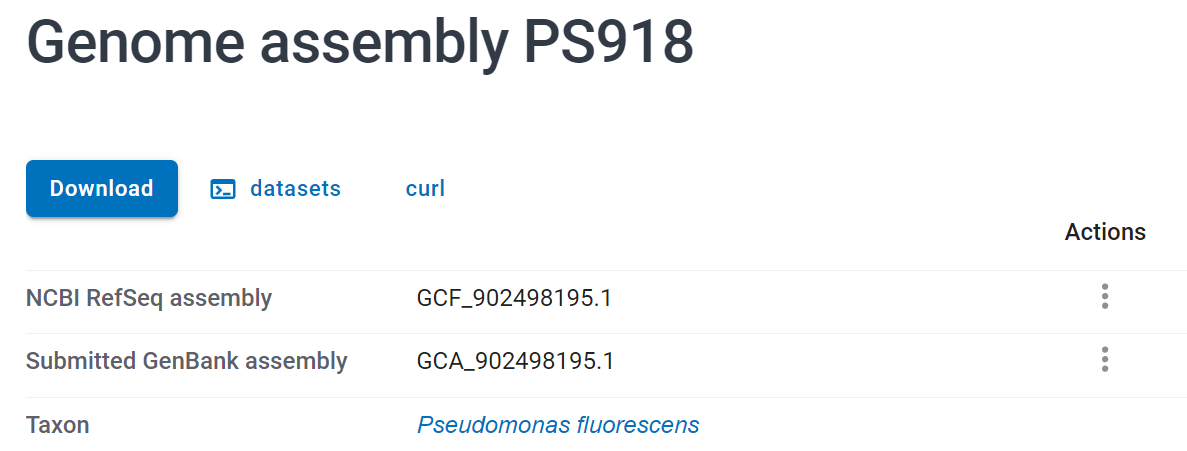
El cual al buscar mas infromacion de este organismo, nos proporciona información como esta



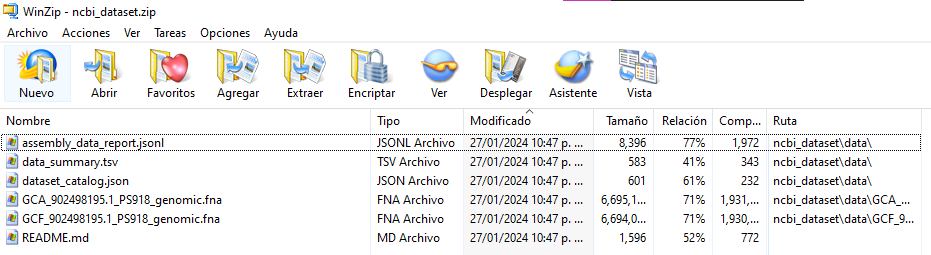
Y que al buscar en el NCBI Assembly Accession: GCA\_902498195.1, nos redirige a la página de NCBI



Que, al escoger la descarga por dos caminos, podemos escoger solo algo de información especializada o una carta de información extra. Para ello podemos ver el botón de **Download.**



El cual te da información como esta, en la que también incluye la secuencia del genoma en \*genomic.fna\*.



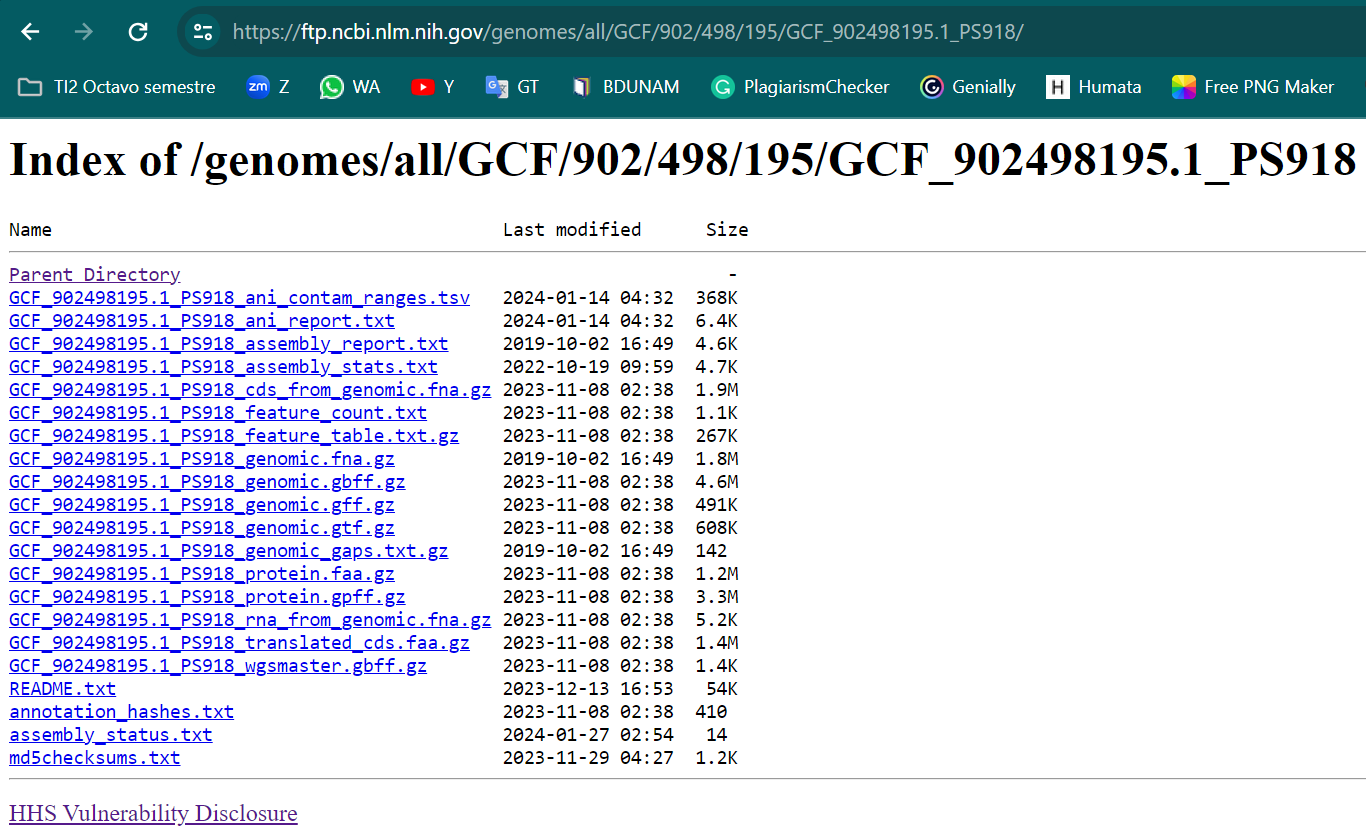
Y la otra forma es seguir esta estructura con el URL especializado de NCBI

“**https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/123/456/789/GCF\_123456789.1\_cepa/**”

El cual te da la dirección completa para encontrar toda la información acerca de esta especie y en este individuo secuenciado, para ello podemos ver el link real:

<https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/902/498/195/GCF_902498195.1_PS918/>

Y nos da toda esta información:



Y para ello ahí mismo podemos encontrar el mismo archivo FNA, pero comprimido “genomic.fna.gz”. Así, con esta disposición para poder descargar los archivos, solo aplicamos algún protocolo como rsync o wget para descargar los archivos respectivos en el cluster y guardarlos.

A continuación, para ello recopilaremos las accesiones a los genomas de cada uno de los organismos que son de interés para aplicar la descarga en un script de forma automática y en lista. Creando una carpeta para solamente las secuencias.

Fecha 📅 28 de enero del 2024

Preguntar a Luis

* ¿Puedo descargar muchos archivos de un solo golpe? **(Intentar y si no, mandar ticket)**
* ¿Hay alguna forma de hacer que al correr un script lo pueda dejar corriendo y me pueda salir del cluster sin ningún problema? (NOHUP) **(TIENE SU METODO, SE PUEDE VER EN LA PAGINA QUE EL COMPARTIO)**
* ¿Para los programas como se puede hacer la solicitud de la instalación? **(POR TICKETS)**

Fecha 📅 29 de enero del 2024

Se empieza el desarrollo de un script

Para esto exploramos comandos y formas de hacer script, de la pagina <https://sites.google.com/a/ciad.mx/bioinformatica/home/genomica/obtencin-de-genomas>

Par lo cual se hará la continuación de la bitácora en Github (NO.1CODE) [[link](https://github.com/Marcos0Ramirez/Pseudomnas_Bitacora/blob/main/Desarrollo/bitacora_desarrollo.md)]

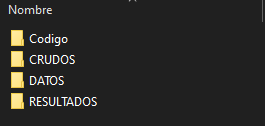
Para eso ahora entramos a la pagina IMG databases, para descargar la tabla de datos de las Pseudomonas patógenas que fueron seleccionadas manualmente y además se les agrego la columna con las accesiones de GenBank. Se agrego el archivo a Github en la carpeta [CRUDO](https://github.com/Marcos0Ramirez/Pseudomnas_Bitacora/tree/main/CRUDO), el nombre del archivo es IMG\_data\_PseudoPahogen.txt.

Por otra parte, el archivo prokaryotes.txt, no se puede añadir a Github, por espacio. Pero se puede descargar del comando que aparece en (NO.1CODE) [[link](https://github.com/Marcos0Ramirez/Pseudomnas_Bitacora/blob/main/Desarrollo/bitacora_desarrollo.md)].

Ya con los datos de las accesiones que se quieren agregar de los organismos patógenos, podemos crear un archivo plano con una lista de accesiones en una sola columna. Asi para que después se agregue un archivo con un script que ejecuta y crea otro archivo con el nombre y el ftp, asi para después agarrar el URL del ftp:// y convertirlo en https:// y después descargar todos los genomas. Para ello nos pasamos a (NO.2CODE) [[link](https://github.com/Marcos0Ramirez/Pseudomnas_Bitacora/blob/main/Desarrollo/bitacora_desarrollo.md)] y creamos una carpeta llamada código y ahí depositamos el script y el archivo “Accessions”

Fecha 📅 30 de enero del 2023

Ahora para ello en la documentación <https://congresos.nnb.unam.mx/TIB2014/sites/default/files/TIB2014/manual_awk.pdf> encontramos un código, el cual se implemento **awk '/GCA\_025447535.1/ { print $0 }' ../proka\_nombre\_accesion\_ftp.txt**, esto con el fin de saber si extrae toda la fila de la accession. Y si lo hace, para ello es implementa en (NO.3CODE) [[link](https://github.com/Marcos0Ramirez/Pseudomnas_Bitacora/blob/main/Desarrollo/bitacora_desarrollo.md)]. Antes mencionar que cambiamos una accesión que tenia GCF por GCA, para que el código hiciese la búsqueda. Así como crear 4 carpetas CRUDOS (Datos descargados y sin modificar), Codigo (donde van los scripts), DATOS (Los datos a usar por los scripts), RESULTADOS (Donde van los resultados).



Y ahora que se quieran extraer los archivos, dentro del script modificado, se debe agregar la forma de descargar el archivo el wget, cambiar la ruta GCA por GCF y finalmente, meter el archivo final que se quiere agregar, copiando la ruta final y agregándole con otro slash, genomic.fna.gz y listo.

Fecha 📅 31 de enero del 2024

¿Cuántas tienen información completa de todo?

* **Pero sobre todo las tablas.**
* Comparar en una tabla, cuantas son de entre cada región, entre hospederos.
* Gráficos para ver el resumen de los datos
* Diferencias entre los individuos (estadísticas genómicas) y cada especie

Por consiguiente, se estuvo explorando el funcionamiento de for, iterado con un While y poder extraer la informacion sin encabezados

Con ello se exploro las paginas webs

* While <https://www.ehu.eus/ehusfera/hpc/2013/11/18/comando-while/>
* Awk (impression limitada) <https://es.stackoverflow.com/questions/397744/c%C3%B3mo-imprimir-un-rango-de-l%C3%ADneas-de-un-fichero>
* For <https://www.hostinger.mx/tutoriales/bash-for-loop-guia-ejemplos>

Y con ello determinar los comandos necesarios para poder hacer la extracción de los datos como en (NO.4CODE) [[link](https://github.com/Marcos0Ramirez/Pseudomnas_Bitacora/blob/main/Desarrollo/bitacora_desarrollo.md)].

Fecha 📅 1 de febrero de 2024

Para este punto se piensa en como se van a cambiar los datos de FTP a HTTP y hacer la descarga con WGET.

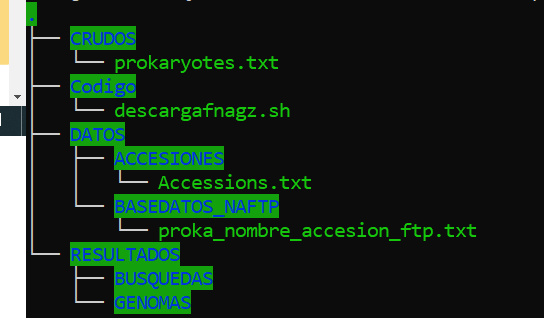
Para ello nos dirigimos a (NO.5CODE) [[link](https://github.com/Marcos0Ramirez/Pseudomnas_Bitacora/blob/main/Desarrollo/bitacora_desarrollo.md)].

Como he estado modificando los datos y sin querer elimine los datos de la carpeta DATOS, voy a modificar las carpetas y ademas de eso, voy a quitar ciertos permisos que son necesarios para evitar la eliminación de la información que sea necesaria.

Se hizo la consulta de como usar las expresiones regulares desde ChatGPT:

* sed: <https://www.youtube.com/watch?v=BE4TsjFIVmM&ab_channel=asteriscom%C3%A1s>, <https://www.infor.uva.es/~mluisa/talf/docs/aula/A2.pdf>, <https://congresos.nnb.unam.mx/TIB2014/sites/default/files/TIB2014/manual_awk.pdf>

Así es la nueva configuración de las carpetas



El código, funciona perfecto, solo que hace falta crear uno de reportes para que diga cuales se descargaron y cuales no, y además de crear otra carpeta de reportes.

Fecha 📅 2 de febrero de 2024

Finalmente, para el código, se agregará al script la forma de que se pueda crear un reporte de las accesiones que no se pudieron descargar. Y además de los registros que hace, cuando empieza a trabajar.

Creamos la carpeta REPORTES y de ahí se añaden los reportes con nombre de la fecha del día y hora, así como el nombre base.

Para el funcionamiento del programa se necesita la función

* locate

Fecha 📅 6 de febrero de 2024

Con el día de hoy se exportaron los datos de IMG databases, para hacer un resumen con los datos. El sistema usado para el análisis es con R.

Fecha 📅 10 de febrero de 2024

Se analiza la base de datos descargada de IMG sobre las pseudomonas, con todas las columnas y datos regristrados. Para este punto, se resolvió la situación de que no se podía abrir todas las filas en R y la tabla no carga, porque tenía una columna extra que no tenía nombre.

Asi, que ahora se encuentra que hay datos que no aparecen en el registro de la tabla, por ejemplo, la fila 13 al 22, si aparece al cargarlo en Excel y al buscarlo directamente en el TXT, pero no en Rstudio.

Pero al contar el numero de columnas de por ejemplo la fila No. 13, tiene el mismo numero de columnas que en el encabezado de cada columna de la tabla. Por lo que aun no puedo determinar, la razón por la que no pueda aparecer en Rstudio.

Al parecer, puede ser que se necesite cambiar los ceros por un campo vacío, debido a que se pueda encontrar números donde van strings o strings en número, pero para acercarnos a la aparición de esos renglones en Excel, puedan aparecer en Rstudio.

Dentro del editor de texto Sublime Text 3:

1. Seleccionamos los 0´s que están solos en un campo por aparte del texto
   1. \t0
2. Y lo cambiamos por
   1. \t

Y así tenemos el mismo numero de columnas (205) ósea 204 tabs.

El resultado parece ser el mismo, me parece peculiar que los datos si cambiaron, los 0 si desaparecieron, incluyendo de la fila 1916 ya no tiene datos. Pero al observar la fila 1916, en el documento de texto si tiene datos, pero lo que es en el Rstudio, no tiene.

Por el momento se seguirá determinando, cuales son los datos que si aparecen en Rstudio y cuales no y ver si hay algún tipo de patrón.

El problema de lectura fue solucionado. La función **read.table** no es la mas optima para abrir un archivo tabulado y de texto, porque se encontraron muchos problemas a la hora de abrir, de tal forma que no se encontraban la gran cantidad de datos, no todos tenían el mismo número de columna y no se podía continuar con nada. Pero ahora se usó la función **read.csv** y el archivo se abrió por completo, con todos los registros vistos desde Excel y Sublime, por tanto, ahora se puede proceder a trabajar con los resúmenes de los datos.

Fecha 📅 11 de febrero de 2024

En este dia, se averiguo que los paquetes de R estuviesen bien y actualizados, ademas de concretar algunos comandos básicos. Para manipular la tabla. Para la cual no se podían procesar los datos de manera correcta debido a que no se habían concretado bien los comandos.

Fecha 📅 12 de febrero de 2024

Se logro separar los datos mas relevantes para analizar en primera instancia, manejando estos tipos de datos

[1] "taxon\_oid" "NCBI.Assembly.Accession" "Species"

[4] "Ecosystem" "Ecosystem.Category" "Ecosystem.Subtype"

[7] "Ecosystem.Type" "Specific.Ecosystem" "Phenotype"

[10] "Relevance" "Temperature.Range" "Geographic.Location"

[13] "Isolation.Country" "Latitude" "Longitude"

[16] "Host.Name" "Genome.Size.....assembled" "GC.Count.....assembled"

Con lo cual nos permitió establecer una tabla con el fin de determinar cuantos individuos hay por especie y sus respectivas accesiones.

especie cantidad

1 Pseudomonas sp. MYb13 1

2 Pseudomonas protegens 34

3 Pseudomonas chlororaphis 75

4 Pseudomonas aeruginosa 975

5 Pseudomonas toyotomiensis 3

6 Pseudomonas daroniae 4

accessiones

1 GCA\_002980075.1

2 GCA\_900560965.1 GCA\_002006545.1 GCA\_001269485.1 GCA\_002901745.1 GCA\_001269465.1 GCA\_001682945.1 GCA\_003731885.1 GCA\_001269475.1 GCA\_001904985.1 GCA\_001269495.1 GCA\_000012265 GCA\_002208745.2 GCA\_001874665.1 GCA\_017923275.1 GCA\_003731825.1 GCA\_003205275.1 GCA\_003731835.1 GCA\_003205455.1 GCA\_001906625.1 GCA\_000397205.1 GCA\_002899905.1 GCA\_001906605.1 GCA\_003732485.1 GCA\_031954405.1 GCA\_003731865.1 GCA\_001906615.1 GCA\_001682935.1

.

.

.

Fecha 📅 13 de febrero de 2024

Solo se usó el código anterior para determinar de donde fueron extraídos, cuantos de ellos pertenecen a cada categoría y que individuos pertenecen a cada una de estas extracciones.

especie cantidad\_database

1 Nematoda 23

2 Roots 255

3 Roots 255

4 Digestive system 30

5 Thermal springs 2

6 Unclassified 1001

especies

1 Pseudomonas sp. MYb13 Pseudomonas brenneri Pseudomonas lurida Pseudomonas cedrina Pseudomonas putida Pseudomonas sp. MYb115 Pseudomonas protegens Pseudomonas sp. JUb52 Pseudomonas helmanticensis Pseudomonas sp. MYb185 Pseudomonas sp. MYb193 Pseudomonas mendocina Pseudomonas sp. MYb3 Pseudomonas poae Pseudomonas trivialis Pseudomonas sp. MYb60 Pseudomonas psychrophila Pseudomonas sp. MYb2 Pseudomonas sp. MYb187

.

.

.

Fecha 📅 14 de febrero de 2024

Por el día de hoy eliminaran de la tabla por especies, las accesiones y se dejara solo la cantidad de individuos por especies y se verá porque hay una repetición de root en la segunda tabla, así como eliminar el nombre de los individuos de quieres fueron extraídos. Y, por último agregarlos a R Markdown nuevo con el fin de mantener código final. Y solo listo para usar y expresar los resultados sin detalles extras, mas que con una explicación amplia.

Revisando a detalle la tabla con el número de individuos por especie, los números y especies se repiten, podemos considerar que en la tabla existe los unclassified y aparece dos veces con exactamente el mismo numero de individuos. Así, revisando la tabla y revisando el numero total de Pseudomonas, encontramos que hay 3868 genomas cargado, para esto. Lo que se hará, es descargar de nuevo toda la base de datos, con todas las columnas para trabajar con la nueva tabla. Pueda ser que se haya duplicado o hecho algo por accidente en el manejo de la tabla o descarga.

**Ahora que descargué de nuevo los datos y los convertí en un txt, pude ver que ya no se repetían los datos y que por ende si uno repite el guardado, se concatenan y no se borran los que estaban.**

Fecha 📅 17 de febrero de 2024

Se repitió el conteo, con respecto a cada una de las columnas que se tomaron inicialmente. Estas son

"Ecosystem.Category" "Ecosystem.Subtype" "Specific.Ecosystem" "Phenotype" "Relevance”

Para entender que más se pueden hacer con los datos, se recopilaron 3 artículos para entender la metodología, además del contexto de las Pseudomonas.

En este punto, se desea, conocer mediante estas preguntas:

¿Cuáles son los datos que usaron?

¿Cuál son los programas que usaron?

¿Por qué lo realizaron de determinada forma y con determinados materiales?

**Artículos**

1. Cesbron, S., Briand, M., Essakhi, S., Gironde, S., Boureau, T., Manceau, C., Fischer-Le Saux, M., & Jacques, M. A. (2015). Comparative Genomics of Pathogenic and Nonpathogenic Strains of Xanthomonas arboricola Unveil Molecular and Evolutionary Events Linked to Pathoadaptation. Frontiers in plant science, 6, 1126. <https://doi.org/10.3389/fpls.2015.01126>
2. Characterization, Pathogenicity, Phylogeny, and Comparative Genomic Analysis of Pseudomonas tolaasii Strains Isolated from Various Mushrooms in China
3. Comparative genomics of Pseudomonas syringae reveals convergent gene gain and loss associated with specialization onto cherry (Prunus avium)
4. **Comparative Genomics of Pathogenic and Nonpathogenic Strains of Xanthomonas arboricola Unveil Molecular and Evolutionary Events Linked to Pathoadaptation**
5. Por resumen, se basaron en diferenciar los que son patogénicos y los no patogénicos por el fenotipo que exhibía la planta por la enfermedad de “bacteria blight” y “VOC”.
6. Compararon los WGSs y encontraron que las cepas patógenas poseen más elementos movibles que las cepas no patógenas.
7. CFBP 7179 posee el elemento integrativo y conjugativo (ICE)
8. Efectores de tipo tres, son mucho más en patógenos que en no patógenos.
9. CFBP 7634 carece de los genes codificantes de la secreción tipo 3
10. CFBP 2528 sistema flagelar, aparece incompleto y no funcional en patógenos de esta cepa.

Alcance a leer, lo que es la introducción, pero me falta leer a metodología…

Fecha 📅 22 de febrero de 2024 a Fecha 📅 26 de febrero de 2024

Lecturas de artículos para genera l introducción de la tesina. Tarea de Ciencias Agrogenomicas. Así como para explorar diferentes rubros que se conocen en Pseudomonas.

Realizado en un Google Docs, llamado “Tesina Pseudomonas”

Fecha 📅 26 de febrero de 2024 a Fecha 📅 27 de febrero de 2024

Se retoma el análisis y resumen de los datos descargados de IMG databases. Para ello continuamos trabajando con el proyecto, usando el archivo Rmarkdown, “FinalPseudomonas\_code.Rmd”, el cual fue creado el 14 de febrero del 2024.

Hasta el momento se hizo un resumen, tanto de cuantos genomas se encuentran por especie, así como de cuantas especies individuos de las especies encontradas, se ve por

|  |  |
| --- | --- |
| No genomas por especie encontrados | Numero de genomas de los cuales tiene registro en determinados hospederos |
|  |  |
| No. Genomas encontrados en categoría de ecosistema | No. Genomas encontrados por subtipo de ecosistema |
|  |  |
| No. Genomas encontrados en ecosistema especifico | No de genomas encontrados por fenotipo |
|  |  |
| No de genomas encontrados por relevancia | No de genomas por locación geografica |
|  |  |
| No de genomas por país aislado | No de genomas por nombre de hospedero |
|  |  |

Fecha 📅 28 de febrero de 2024

El día de hoy se asignó la tarea de tomar de nuevo, una descarga de los datos actualizada, proveniente de la base de datos IMG, para descargar las nuevas accesiones y también categorizar los datos por hospedero y no hospedero, así como subtipos.

Fecha 📅 29 de febrero de 2024 a Fecha 📅 1 de marzo de 2024

Lectura de artículos para realización de tesina, avance en el archivo Tesina Pseudomonas.

Fecha 📅 2 de marzo de 2024

Avance de código para purificar los datos. Para el Rmarkdown llamado “IMGpseudomonas.Rmd” para comprobar el funcionamiento del código, con algunas pruebas y FinalPseudomonas\_code.Rmd para colocar el código final.

Para este caso IMGpseudomonas.Rmd, se quiere comprobar como un vector se puede añadir a una base de datos y agregarle el nombre con respecto a la columna por la cual fueron extraídos los datos.

Conocer la función rbind, ayuda a que un vector pueda ser concatenado en distintas columnas. Por lo que ahora se modificara para que los datos se concatenen al final por columna y permita que puedan ser comparados después.

Ahora que ya se tiene el dataframe, se quiere identificar cuales con los datos que no tengan las accesiones de NCBI

Fecha 📅 3 de marzo de 2024

Continuamos con el trabajo en FinalPseudomonas\_code.Rmd, para encontrar aquellos datos que puedan representar y no se encuentren en otros campos.

Para este bloque se tomara a comparar con todos y cada uno ¿Cómo?

Se comparar en base a uno contra todos y el siguiente, contra todos menos con el primero que ya se comparo

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Fila 1 | Biotic. Relationships | Habitat | Isolation | Phenotype | Relevance | Host.Name | Diseases | Ecosystem | Ecosystem. Category | Ecosystem. Subtype | Ecosystem. type | Specific. Ecosystem |
| Biotic. Relationships |  |  |  |  |  |  |  |  |  |  |  |  |
| Habitat |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  | | | | | | | | | | | |
| Ecosystem. type |  |  |  |  |  |  |  |  |  |  |  |  |
| Specific. Ecosystem |  |  |  |  |  |  |  |  |  |  |  |  |

1. Considerando que la primera columna se compara con todas y todas salen F, quiere decir que es el único y podemos seguir con la siguiente fila
2. Si en todo caso, se compara que no todos son F, y que hay uno que es V, se pasa, hasta el que es V y se compara con todos los demás que sean F o V
   1. Si todos los restantes son F, entonces lo consideramos como único y continuamos con la siguiente fila.
   2. Si hay uno otro que sea V, entonces lo ponemos en otra categoría para clasificación
      1. Esto quiere decir que pondremos las filas en los que solo tengan una V, otras que solo sean dos V y asi hasta ver la posibilidad de que todas sean V o no haya ninguna V.

¡Pensándolo, bien, sería mejor utilizar los datos originales, debido a que hay datos perdidos que dicen unclassified y que por tanto tampoco se podrían usar para clasificar los datos, si en todo caso es el único dato presenten en todas las columnas de al menos una!

**SE MIGRARON LAS LINEAS ANTERIORES 368 – 491 DE FinalPseudomonas\_code.Rmd a IMGPseudomonas.Rmd y ahora se trabaja en IMG…, para solo dejar el código final en el script Final…**

Continua…

Antes, debemos considerar que si se quiere filtrar aquellos que no tengan datos o no estén clasificados, se debe saber cómo se distribuyen los datos, de esto tiene que ver las tablas que se hicieron anteriormente, para explorar la distribución de los datso de cada una de las columnas.

Para ello

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Biotic. Relationships | Habitat | Isolation | Phenotype | Relevance | Host.Name | Diseases | Ecosystem | Ecosystem. Category | Ecosystem. Subtype | Ecosystem. type | Specific. Ecosystem |
|  |  |  |  |  |  |  |  |  |  |  |  |

Fecha 📅 4 de marzo de 2024 - Fecha 📅 7 de marzo de 2024

Se modifico el codigo para hacer la categorización de la información, de la cual todo se encuentra en el archivo IMGPseudomonas.Rmd

Fecha 📅 8 de marzo de 2024

Se descargaron los genomas posibles, con las anotaciones hechas en IMG directamente desde la página de IMG y avance de lectura para tesina.

<https://genome.jgi.doe.gov/portal/ext-api/downloads/img/294448-5/status>

Fecha 📅 9 de marzo de 2024 - Fecha 📅 12 de marzo de 2024

Se clasificaron los datos por hospederos y no hospederos, la continuación de la bitácora se encuentra en el Rmarkdown “IMGPseudomonas.Rmd”, para finalmente, tener todo en limpio en “FinalPseudomonas\_code.Rmd”. Así mismo para el día 10 de marzo del 2024, llego la solicitud de descarga para las secuencias de Pseudomonas. Y para este 12 se comenzó a ingresar la carpeta de los genomas, dentro del cluster.

Fecha 📅 13 de marzo de 2024

Finalmente, de la descarga, a la hora de descargar los datos de los genomas, me mando algunos genomas que no se pudieron descargar y sus accesiones de IMG databases:

Download request completed.

Download (22GB, tar file)

The data will be stored on the server for 14 days.

Data for the following datasets are not available:

8084820202, 8085341477, 8084395441, 8085688997, 8085347597, 8084092649, 8084807198, 8085085956, 8019775933, 637000224, 8085016530, 8083583212, 8085092110, 8084839925, 8084413206, 8085098197, 8084767992, 8086259606, 8084850413, 8084102989, 8085237721, 8084400239, 8084628194, 8084693108, 8084813701, 8084762373, 8083775200, 8084856734, 8084097705, 8084833194, 8023655410, 8084826874

Fecha 📅 15 de marzo de 2024

Se analizó la descarga de los genomas dentro del cluster DNA, considerando la cantidad total de genomas descargados por carpetas y cuantos “.FAA” son, si faltan, entonces se procede a hacer la anotación de los pocos genomas que se encuentren sin anotación. Así para proceder a la clusterización y clasificación de los genomas de descarga

Aplicamos los comandos para extraer todos los genomas:

1. Creamos carpeta “PSEUDOMONAS\_GENOMAS”
   1. mkdir PSEUDOMONAS\_GENOMAS
2. extraemos los genomas
   1. tar -xvf img\_data\_294448-5.tar -C ./PSEUDOMONAS\_GENOMAS
3. Descomprimimos los archivos
   1. for descom in \*.tar.gz; do tar -xzf "$descom"; done
4. Pero como se descomprimen y se genera la carpeta y se mantiene los genomas comprimidos, los movemos
   1. mkdir ../COMPRESS\_PSEUDOMONAS\_GENOMAS
   2. mv ./ PSEUDOMONAS\_GENOMAS /\*tar.gz ../COMPRESS\_PSEUDOMONAS\_GENOMAS/
5. Penúltimo contamos el número de carpetas que hay en PSEUDOMONAS\_GENOMAS
   1. ls | wc -l
6. Contamos el numero de carpetas que tienen la anotación “.faa”, si coinciden con los numero, entonces todos están anotados
   1. ls -l \* | grep ".faa" | wc -l

Es asi que ambos indican 3894 carpetas y 3894 “.FAA” respectivamente.

Así se leyó un poco de cada programa CD-HIT y VSEARCH.

Fecha 📅 16 de marzo de 2024

Se retomo lectura sobre CD-HIT y VSEARCH, así como bibliografía para tesina.

Fecha 📅 16 de marzo de 2024

Clasificación de los datos para tipos de hospederos, así como distribución de tablas finales con la clasificación:

|  |  |  |
| --- | --- | --- |
| Distribución | Dataframes con filas clasificadas | A extracción de tabla original |
| Hospedero | ABC\_ConHospedero | PseudoAll |
| No hospedero | ABC\_NoHospedero |
| Sin Clasificación | ABC\_SinData11\_sinclashost |

Al ver que dentro de la tabla original hay dos columnas que parece ser que tienen el mismo formato e incluso los mismo datos por fila:

1. taxon\_oid
2. IMG.Genome.ID

Y tomando de referencia algunas accesiones de genomas que no se pudieron descargar. Se pudo corroborar, uno: que los genomas se identifican por estas dos columnas mencionadas anteriormente. Dos: Que con la prueba de algunas accesiones se puedo comprobar que las columnas son lo mismo, se usó el comando

PseudoAll[PseudoAll$taxon\_oid == 8085688997,] == PseudoAll[PseudoAll$IMG.Genome.ID == 8085688997,]

Y con este comando, se determinó que todas las columnas e incluso fila coincida exactamente, por lo tanto, se puede utilizar cualquiera de estas dos columnas para la descarga y clasificación de los genomas por accesiones de taxon o id IMG.

Fecha 📅 17 de marzo de 2024

Ya que inicialmente tomamos estas columnas para clasificar los genomas

|  |  |
| --- | --- |
| Columna Nombre Clasificatorio | No. Columna de tabla original |
| Ecosystem | 63 |
| Ecosystem.Category | 64 |
| Ecosystem.Subtype | 65 |
| Ecosystem.Type | 66 |
| Specific.Ecosystem | 67 |
| Host.Name | 109 |
| Habitat | 90 |
| Isolation | 91 |

Por último se usara de las columnas Phenotype, y algunos que son de Relevance e indican que son patógenos de algún grupo.

Fecha 📅 18 de marzo de 2024

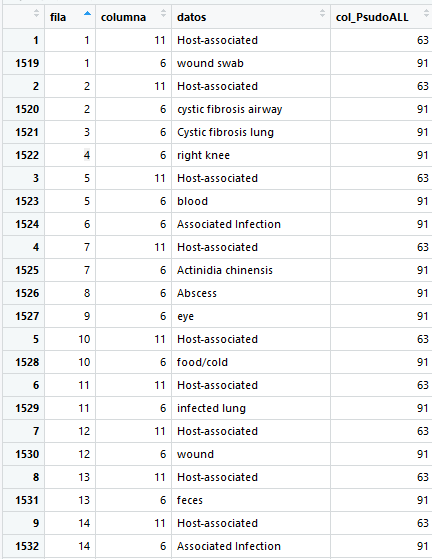
Después de crear las tablas filtradas por filas y en acorde a si es hospedero o no, o sin clasificación, se determinó que la columna Phenotype, en la tabla con registro sin clasificación no puede ser útil, porque no hay ningún registro útil, más que solo campos vacíos.

Finalmente queda la columna Relevance, la cual tiene, de manera manual será determinada:

“Empty”, tampoco cuanta con datos disponibles, por lo que podemos concluir que se tienen todos los datos clasificados en la primera categoría.

Antes de continuar con la clasificación, se encontró un error, con respecto a la clasificación primaria, hay datos que no corresponden que se encuentran en las tablas Con y Sin hospedero, se revisara todo el código, para ver la repetición de ellos y la no correspondencia con respecto al tipo de dato.

Todo empieza con que en la tabla ABC\_ConHospedero tiene filas repetidas, si lo ordenamos por filas aparecen repetidos



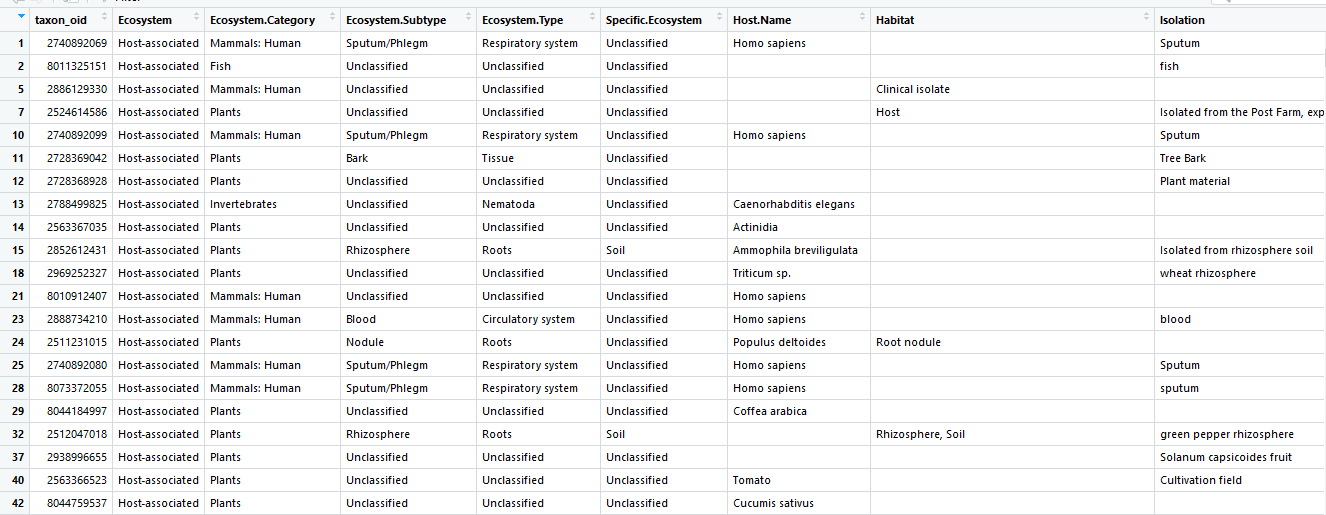
Fue claro, por que en la tabla final “tipo\_host\_sinclasificar\_PseudoAll”

tipo\_host\_sinclasificar\_PseudoAll <- Host\_PseudoAll[,c(1, 63, 64,

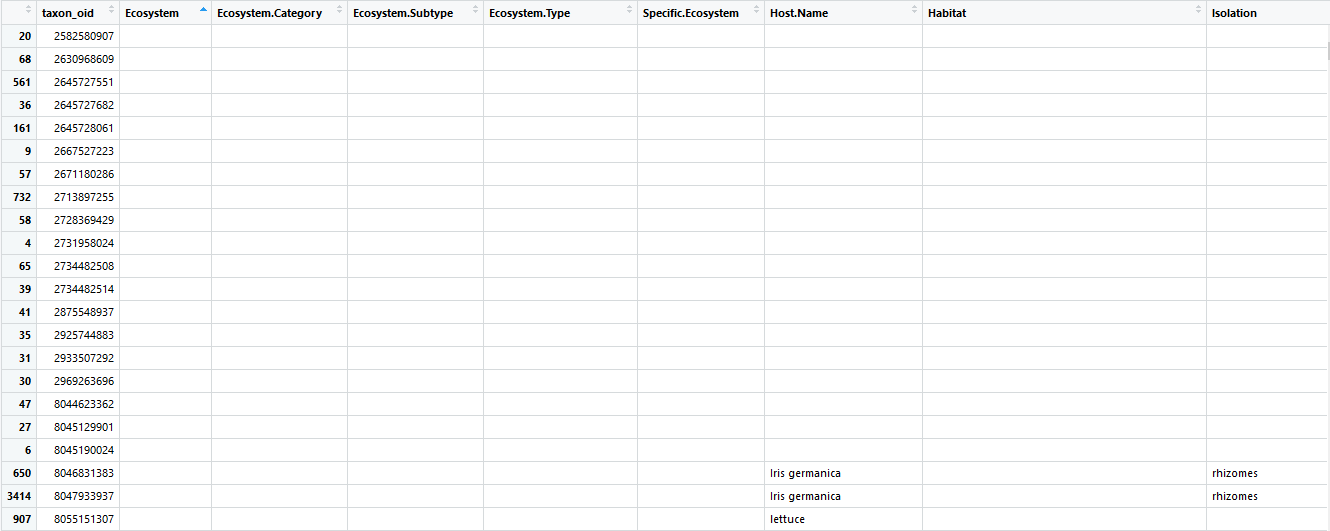
65, 66, 67,

109, 90, 91)]

al organizar los datos para ver si hay irregularidades, se encontró que al ver las tablas ordenadas por posición del dato, no parecía haber algún error, hasta que se ordeno por Ecosystem y se encontró que había filas que no correspondían.



Como vemos faltan datos



Y luego vemos en la tabla ABC\_ConHospedero y vemos que se repiten algunas filas, contemplando que el código que filtra los datos que pasan de la tabla en la que contiene las filas que no están clasificadas a alguna de las clasificadas, pasa sin discriminación alguna, mas que el mero hecho de que las filas ya están clasificadas. Por tanto, se revisará el código desde la parte en la que se crearon las tablas ABC. Para ver luego la que filtra de una tabla a otra.

Después de revisar detenidamente los códigos, se subraya el potencial error en el Rmarkdown “FinalPseudomonas\_code.Rmd” de las líneas 517 a 613. Para ello estos códigos los pasamos para el final del Rmarkdown de “IMGPseudomonas.Rmd”. Y lo corregimos para modificar y ver el potencial error, se encontrará en las líneas 1007 a1103

Finalmente se termino de modificar y se pudo mejorar el código, por lo que ya se comprobó que no se repiten de nuevo las filas.

Por lo que si corremos todo el script de Final…, sin los cambios, al verificar, usando los comando, aparece esto

```{r}

print(paste(length(ABC\_NoHospedero$fila), length(unique(ABC\_NoHospedero$fila))))

print(paste(length(ABC\_ConHospedero$fila), length(unique(ABC\_ConHospedero$fila))))

unique(c(1,2,3,1))

```

[1] "877 866"

[1] "1592 1552"

[1] 1 2 3

Pero si lo cambias por el código final, corregido, no hay cambios en los números, lo que indica que no hay números repetidos.

[1] "877 877"

[1] "1592 1592"

[1] 1 2 3

Así constamos de que todo está correcto y sin repeticiones.

Líneas nuevas en Final… de 614 a 645

Fecha 📅 19 de marzo de 2024

El día de hoy se continua con la clasificación de los tipos de hospederos

tipo\_host\_sinclasificar\_PseudoAll <- Host\_PseudoAll[,c(1, 63, 64,

65, 66, 67,

109, 90, 91)]

Y podemos ver que no faltan datos en la tabla “tipo\_host\_sinclasificar\_PseudoAll”, Para clasificar todos los que vienen de plantas, solo se dio una revisada manual, por medio del buscador que tienen las tablas en Rstudio y se determinó que la gran mayoría de salidas vienen de Ecosystem.Category y Habitat.

Con la tabla final, se categorizaron los datos de la tabla lista\_otros de manera manual en un lista de las posiciones y transferir a las tablas respectivas.

Fecha 📅 20 de marzo de 2024

1. Clasificación de datos
2. Plantas terrestres
3. Plantas acuáticas
4. Hongos
5. Mamíferos
6. Humanos

# Eliminar de los datos, los que los genomas que no se descargaron.

Y en CD HIT y Vsearch se corre todo junto.

Para determinar que genes son importantes para cada grupo

Plantas terrestres vs mamíferos vs no hospederos

Primero probar con pocos genomas y luego hacer script para usar todos los genomas.

Para ambos programas pide el umbral de considera una proteína.

60% de identidad en CD HIT es mas directo y en vsearch, ver cómo poner este valor.

Ya con el programa instalado de CD-HIT, se procedio a ejecutar el comando con un genomas en distintos archivos para el mismo comando,

Por default se intento con el código inicial

**cd-hit -i nr -o nr100 -c 1.00 -n 5 -M 16000 –d 0 -T 8**

pero como no habia ningún archivo con ese nombre se tomaron 10 secuencias descargadas de IMG database y se intento con un solo archivo **cd-hit -i 2505313052.fna -o nr100 -c 1.00 -n 5 -M 800 -d 0 -T 1**

del cual solo se trata de un genoma y por tanto en este exporta dos archivos

|  |  |
| --- | --- |
| nr100 | nr100.clstr |
|  |  |

Por ende no mostraba datos significativos, asi se aplico lo mismo, pero con otro archivo del mismo genoma

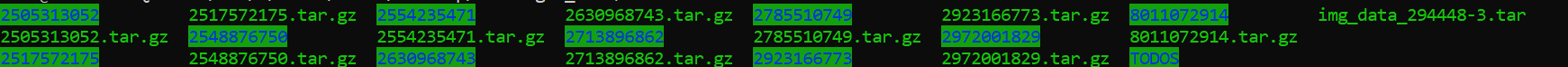
**cd-hit -i 2505313052.genes.fna -o 2505313052genes -c 1.00 -n 5 -M 800 -d 0 -T 1**

con resultados como

|  |  |
| --- | --- |
| 2505313052genes | 2505313052genes.clstr |
|  |  |

Con un resultado similar al anterior.

Pero después se concatenaron de los genomas existentes



Se concatenaron los genomas

**cat 2505313052/2505313052.genes.fna >> allPseudos.genes.fna**

**cat 2554235471/2554235471.genes.fna >> allPseudos.genes.fna**

dentro de la carpeta TODOS/ y se uso para aplicar el siguiente comando

**cd-hit -i allPseudos.genes.fna -o allgenes -c 1.00 -n 5 -M 800 -d 0 -T 1**

con resultados similares

|  |  |
| --- | --- |
| allgenes | allgenes.clstr |
|  |  |

Por tanto se procedió a aplicarlo con genomas concatenándolos en un archivo

**cat ../2505313052/2505313052.fna >> PSUEDOS.fna**

**cat ../2517572175/2517572175.fna >> PSUEDOS.fna**

Y se cambio la identidad mínima de 60

**cd-hit -i PSUEDOS.fna -o genomePSUEDO -c 0.60 -n 5 -M 800 -d 0 -T 1**

Pero la compilacion de CD-HIT no estaba optimizada para correr este comando, fue asi hasta jugar a cambiarle los parámetros

**718 cd-hit -i PSUEDOS.fna -o genomePSUEDO -c 0.60 -n 4 -M 1400 -d 0 -T 1**

**719 cd-hit -i PSUEDOS.fna -o genomePSUEDO -c 0.60 -n 4 -M 1400 -d 0 -T 4**

**720 cd-hit -i PSUEDOS.fna -o genomePSUEDO -c 0.60 -n 4 -M 1400 -d 0 -T 2**

**721 cd-hit -i PSUEDOS.fna -o genomePSUEDO -c 0.70 -n 4 -M 1400 -d 0 -T 1**

**722 cd-hit -i PSUEDOS.fna -o genomePSUEDO -c 0.90 -n 5 -M 1400 -d 0 -T 1**

**723 cd-hit -i PSUEDOS.fna -o genomePSUEDO -c 0.90 -n 5 -M 800 -d 0 -T 1**

**724 cd-hit -i PSUEDOS.fna -o genomePSUEDO -c 0.90 -n 2 -M 800 -d 0 -T 1**

**725 cd-hit -i PSUEDOS.fna -o genomePSUEDO -c 1.00 -n 5 -M 800 -d 0 -T 1**

**726 cd-hit -i PSUEDOS.fna -o genomePSUEDO -c 1.00 -n 5 -M 1400 -d 0 -T 1**

Pero de ninguna de las anteriores acepto, hasta que regrese con el comando anterior, pero con una identidad del 60%

**cd-hit -i allPseudos.genes.fna -o allgenes2 -c 0.60 -n 4 -M 800 -d 0 -T 1**

Fue así que después de 2 hrs aproximadamente resulto en los archivos siguientes

|  |  |
| --- | --- |
| allgenes2 | allgenes2.clstr |
|  |  |

Mostrando que ahora los archivos los genes si se habían clusterizado, pero para esto se confirmo usando grep si las accesiones de los genes en cada uno de los genomas, se encontraban y si en efecto si resultaba el análisis correctamente.

Ahora solo queda determinar que resulta de usar .FAA y por 4 genomas, asi como determinar si el cluster se refiere a clasificación dentro fragmentos del genoma o si por genes.

Fecha 📅 21 de marzo de 2024

Se procedió a entender que puede pasar al cambiar los valores del script, pero se solucionó una pregunta ¿Cómo se puede agrupar todos los genomas de las Pseudomonas? Si, la solución tiene que ver con concatenar los genomas en uno solo y de ahí correr el script para que los agrupe, solo falta hacer el script para concatenar todos y así el script correcto para hacer la agrupación.

Antes, ayudara a ver la información expedida por el comando, la cual es descrita en Github <https://github.com/Marcos0Ramirez/Pseudomnas_Bitacora/blob/main/CD-HIT/README.md>

Fecha 📅 22 de marzo de 2024

Continua la bitácora en Github en la carpeta CD-HIT, descripción den README.md

Por otra parte se analizo como usar el programa de VSEARCH

<https://github.com/Marcos0Ramirez/Pseudomnas_Bitacora/blob/main/VSEARCH/LENOVO/README.md>

Fecha 📅 23 de marzo de 2024

Clasificación de genomas por los tipos de hospederos.

1. Clasificación de datos
2. Plantas terrestres
3. Plantas acuáticas
4. Hongos
5. Mamíferos
6. Humanos

Fecha 📅 24 de marzo de 2024

Se continuo con la clasificación

Fecha 📅 25 de marzo de 2024

Con las clasificación, se realizo un código sencillo para transferir la informaicon a las tablas correspondeintes, ademas de transformar las tablas

1. Host\_PseudoAll
2. NoHost\_PseudoAll
3. Unclass\_PseudoAll

Ya con los datos se paso el código final al Rmd, Final…

Fecha 📅 1 de abril de 2024

Continuamos con la contruccion del comando a ejecutar para usar VSEARCH

<https://github.com/Marcos0Ramirez/Pseudomnas_Bitacora/blob/main/VSEARCH/LENOVO/README.md>

Se realizo un script para correr el análisis en VSEARCH, asi para concatenar las secuencias de ADN, porque vsearch indicaba que no podía usar las de proteínas y que probablemente saldría mal el análisis. De mas desenlace en <https://github.com/Marcos0Ramirez/Pseudomnas_Bitacora/tree/main/VSEARCH/DNA>

Fecha 📅 3 de abril de 2024

|  |  |
| --- | --- |
| Clasificación para 3923 genomas (aun sin restar los no descargados) | |
| Categoría | Genomas |
| Pseudomonas sin hospedero (ambiente) | 885 |
| Pseudomonas con hospedero | 1592 |
| Hospedero algas | 8 |
| Hospedero animales | 62 |
| Hospederos animales acuáticos | 22 |
| Hospedero ambiente | 4 |
| Hospedero hongo | 28 |
| Hospedero humano | 512 |
| Hospedero insectos | 22 |
| Hospedero plantas | 923 |
| Hospedero plantas acuáticas | 4 |
| NO CLAROS DE HOSPEDEROS | 7 |
| Sin clara clasificación (no info para clasificar) | 1454 |

* Falta clasificar cuales son las plantas acuáticas y cuales son terrestres
* Asi, como los patógenos y los no patógenos
* Agregar todo a la tesina
* Y falta solo definir si hay alguno de no hospedero (ambiente que aclare que son hospederos)
* Restar los que no pudieron ser descargados.

Fecha 📅 4 de abril de 2024

Se comenzó con el bash script para realizar la matriz de la salida de información de CD-HIT

Fecha 📅 5 - 16 de abril de 2024

Se continuo con los scripts de CDHIT y VSEARCH, y ya terminando. <https://github.com/Marcos0Ramirez/Pseudomnas_Bitacora> Se puede ver en las carpetas correspoendientes en las carpetas dentro llamadas como MATRIZ en CDHIT y DNA y LENOVO en VSEARCH.

Fecha 📅 18 de abril de 2024

Ahora se encuentra en operación el script de matriz de CDHIT, Ahora que se vio que VSEARCH no procesa proteínas y que por tanto no es comparable con el análisis que hizo CDHIT con las proteínas, se pretende usar ahora USEARCH y RAFTS3G (<https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-019-2973-4>)

<https://github.com/Marcos0Ramirez/Pseudomnas_Bitacora/tree/main/CD-HIT/MATRIZ/DNA#fecha-18-de-abril-del-2024>

Fecha 📅 19 de abril de 2024

Termino de correr en el cluster VSEARCH <https://github.com/Marcos0Ramirez/Pseudomnas_Bitacora/tree/main/VSEARCH/DNA>

Fecha 📅 20 de abril de 2024 a 24 de abril de 2024

Modificacion en <https://github.com/Marcos0Ramirez/Pseudomnas_Bitacora/blob/main/CD-HIT/MATRIZ/README2.md>

Fecha 📅 27 de abril de 2024 a 29 de abril de 2024

Modificacion en <https://github.com/Marcos0Ramirez/Pseudomnas_Bitacora/blob/main/CD-HIT/MATRIZ/README2.md>

Fecha 📅 1 de mayo de 2024 a 5 de mayo de 2024

Modifiacaciones

<https://github.com/Marcos0Ramirez/Pseudomnas_Bitacora/blob/main/CD-HIT/MATRIZ/README2.md>

<https://github.com/Marcos0Ramirez/Pseudomnas_Bitacora/tree/main/CD-HIT/MATRIZ/DNA>

<https://github.com/Marcos0Ramirez/Pseudomnas_Bitacora/blob/main/CD-HIT/MATRIZ/DNA/ERRORES.md>

Fecha 📅 5 de mayo de 2024 a 12 de mayo de 2024 sin contar 8 y 10

<https://github.com/Marcos0Ramirez/Pseudomnas_Bitacora/blob/main/Random%20Forest/RandomForest_Rstudio.md>

Tambien como se pudo comprobar, dentro del cluster, si están las paqueterías necesarias de numpy, pandas y de https://scikit-learn.org/stable/ , pero no de la versión 3 de Python, sino en la versión 2 de Python.

Para esto se modificara solo, la entrada de Python en el código de bash para hacer la matriz <https://github.com/Marcos0Ramirez/Pseudomnas_Bitacora/tree/main/CD-HIT/MATRIZ>

y también el código para hacer el Randomforest

<https://github.com/Marcos0Ramirez/Pseudomnas_Bitacora/blob/main/Random%20Forest/RandomForest_Rstudio.md>

Para ello, hay que saber que versión de pyhton 2 hay, para aplicarlo localmente y luego verificar que funciona.

Así, viendo que se encuentra Python 2.7.5, podemos verificar que los scripts funcionen correctamente.

Finalmente para entender si puede funcionar el script de bash con Python, lo aplicamos

<https://github.com/Marcos0Ramirez/Pseudomnas_Bitacora/blob/main/CD-HIT/MATRIZ/README2.md>

Por lo visto al tener la versión y solo llamarla si, funciona se puede aplicar.

Así, al momento se usar randomForest podrá ser ocupado con la versión de Python 2.7.5, dentro del cluster.

Asi, empezamos a trabajar en el cluster

<https://github.com/Marcos0Ramirez/Pseudomnas_Bitacora/tree/main/CD-HIT/MATRIZ/DNA>

1. Realizando el código para que use Python 2.7.5 para la matriz

Fecha 📅 15 de mayo de 2024

Se trato de verificar el problema sobre los id extras

<https://github.com/Marcos0Ramirez/Pseudomnas_Bitacora/blob/main/CD-HIT/MATRIZ/DNA/ERRORES.md>

Ya se encontró y se llegó a la conclusión de hacer un pequeño código para que filtre los datos antes de usarlos.

Y también dejamos corriendo CD-HIT con 90% de identidad para aplicar la matriz.

<https://github.com/Marcos0Ramirez/Pseudomnas_Bitacora/tree/main/CD-HIT/DNA>

Y por otra parte realizamos modificación en el código para hacer la matriz

<https://github.com/Marcos0Ramirez/Pseudomnas_Bitacora/blob/main/CD-HIT/MATRIZ/README2.md>

Fecha 📅 16 de mayo de 2024

Se logro terminar la corrección sobre la generación de la matriz.

<https://github.com/Marcos0Ramirez/Pseudomnas_Bitacora/blob/main/CD-HIT/MATRIZ/DNA/ERRORES.md>

Finalmente los adaptamos al código madre

<https://github.com/Marcos0Ramirez/Pseudomnas_Bitacora/blob/main/CD-HIT/MATRIZ/DNA/README.md>

Fecha 📅 17 de mayo de 2024

Se actualizaron los numero de las tablas,

1. agregando los que son patógenos por descripción, removiéndolos de sus respectivas tablas
2. Corrigiendo el numero para que sumen los 2923
3. Eliminando de estos metadatos los 32 genomas que no estaban disponibles
   1. 8084820202, 8085341477, 8084395441, 8085688997, 8085347597, 8084092649, 8084807198, 8085085956, 8019775933, 637000224, 8085016530, 8083583212, 8085092110, 8084839925, 8084413206, 8085098197, 8084767992, 8086259606, 8084850413, 8084102989, 8085237721, 8084400239, 8084628194, 8084693108, 8084813701, 8084762373, 8083775200, 8084856734, 8084097705, 8084833194, 8023655410, 8084826874

Para empezar con el código de FinalPseudomonas\_code.Rmd se eliminaran los taxoid que no fueron descargados.

El total eran 3923 y eliminando los 32 genomas, quedan 3891, pero los que fueron descargados son 3894, por lo que hay 3 genomas que no están en los metadatos y que deben ser explorados.

Asi al meternos en IMG databases

|  |  |
| --- | --- |
| Clasificación para 3894 genomas | |
| Categoría | Genomas |
| Pseudomonas sin hospedero (ambiente) | 885 |
| Pseudomonas con hospedero | 1592 |
| Hospedero algas | 8 |
| Hospedero animales | 62 |
| Hospederos animales acuáticos | 22 |
| Hospedero ambiente | 4 |
| Hospedero hongo | 28 |
| Hospedero humano | 512 |
| Hospedero insectos | 22 |
| Hospedero plantas | 923 |
| Hospedero plantas acuáticas | 4 |
| NO CLAROS DE HOSPEDEROS | 7 |
| Sin clara clasificación (no info para clasificar) | 1454 |
|  |  |

Fecha 📅 18 de mayo de 2024

Finalmente se pudieron clasificar todos lo genomas en la primera categoría, haciendo la limpieza y uso de solo los genomas que si están presentes en los análisis. El nombre del R es Pseudomonas\_Genomeclass.Rmd

|  |  |
| --- | --- |
| Clasificación para 3894 genomas | |
| Categoría | Genomas |
| Pseudomonas sin hospedero (ambiente) | 872 |
| Pseudomonas con hospedero | 1580 |
| Sin clasificación | 1442 |

Fecha 📅 19 de mayo de 2024

El día de hoy se verifico como hacer una matriz de confusión con los datos de prueba, en el script RandomForest\_pagedevelop.py

Al final se desarrolló otro llamado Grafio\_Confusion\_RandomForest.py que esta guardado en el repositorio <https://github.com/Marcos0Ramirez/Pseudomnas_Bitacora/tree/main/Random%20Forest>

Fecha 📅 22 de mayo de 2024

Se corrio el script de Python para generar la matriz de Confusión y también el grafico con los cluster con mayor influencia en la predicción de individuos no hospederos y hospederos <https://github.com/Marcos0Ramirez/Pseudomnas_Bitacora/tree/main/Random%20Forest/DNA>

Por otra parte, se hará la clasificación entre mamífero versus planta en el script de R.

Fecha 📅 23 de mayo de 2024 al 27 de mayo del 2024

Se trabajo en RandomForest modificando el script de análisis de Random Forest, hasta que final quedo funcionando correctamente y con resultados buenos de las predicciones. Aquí los avances finales del script de RandomForest en Python. <https://github.com/Marcos0Ramirez/Pseudomnas_Bitacora/blob/main/Random%20Forest/README.md> y el link con los scripts realizados <https://github.com/Marcos0Ramirez/Pseudomnas_Bitacora/tree/main/Random%20Forest/DNA>

Fecha 📅 28 de mayo de 2024

Presentación final en la ENES León para Agrogenómicas

Fecha 📅 29 de mayo del 2024 al 2 de junio del 2024

Se clasificaron los datos entre Plantas y Mamíferos en Rstudio. Link donde se encuentra el script nuevo y el set de datos <https://github.com/Marcos0Ramirez/Pseudomnas_Bitacora/tree/main/Ranalisis/LIMPIO>